Quantitative Finance, 11:7, 991-1012 (2011)
http://dx.doi.org/10.1080/14697688.2010.539248

Econophysics Review: I. Empirical facts

Anirban Chakraborti[a] and Ioane Muni Toke[b]
Laboratoire de Mathématiques Appliquées aux Systèmes, Ecole Centrale Paris, 92290 Châtenay-Malabry, France
Marco Patriarca[c]
National Institute of Chemical Physics and Biophysics, Rävala 10, 15042 Tallinn, Estonia and
Instituto de Fisica Interdisciplinaire y Sistemas Complejos (CSIC-UIB), E-07122 Palma de Mallorca, Spain
Frédéric Abergel[d]
Laboratoire de Mathématiques Appliquées aux Systèmes, Ecole Centrale Paris, 92290 Châtenay-Malabry, France

(Dated: 3 November 2010)

This article and its companion paper to follow aim at reviewing recent empirical and theoretical developments usually grouped under the term *Econophysics*. Since its name was coined in 1995 by merging the words "Economics" and "Physics", this new interdisciplinary field has grown in various directions: theoretical macroeconomics (wealth distributions), microstructure of financial markets (order book modelling), econometrics of financial bubbles and crashes, etc. We discuss interactions between Physics, Mathematics, Economics and Finance that led to the emergence of Econophysics. Then we present empirical studies revealing statistical properties of financial time series. We begin the presentation with the widely acknowledged "stylized facts" which describe the returns of financial assets – fat tails, volatility clustering, autocorrelation, etc. – and recall that some of these properties are directly linked to the way "time" is taken into account. We continue with the statistical properties observed on order books in financial markets. For the sake of illustrating this review, (nearly) all the stated facts are reproduced using our own high-frequency financial database. Finally, contributions to the study of correlations of assets such as random matrix theory and graph theory are presented. The following companion paper will review models in Econophysics through the point of view of agent-based modelling.

Keywords: Econophysics; Stylized facts ; Financial time series; Correlations; Order book models ; Agent-based models; Wealth distributions; Game Theory; Minority Games; Pareto Law; Entropy maximization; Utility maximization.

PACS Nos.: 05.45.Tp, 02.50.Sk, 05.40.-a, 05.45.Ra, 89.75.Fb

I. INTRODUCTION

What is Econophysics? Fifteen years after the word "Econophysics" was coined by H. E. Stanley by a merging of the words 'Economics' and 'Physics', at an international conference on Statistical Physics held in Kolkata in 1995, this is still a commonly asked question. Many still wonder how theories aimed at explaining the physical world in terms of particles could be applied to understand complex structures, such as those found in the social and economic behaviour of human beings. In fact, physics as a natural science is supposed to be precise or specific; its predictive powers based on the use of a few but universal properties of matter which are sufficient to explain many physical phenomena. But in social sciences, are there analogous precise universal properties known for human beings, who, on the contrary of fundamental particles, are certainly not identical to each other in any respect ? And what little amount of information would be sufficient to infer some of their complex behaviours ? There exists a positive strive in answering these questions. In the 1940's, Majorana had taken scientific interest in financial and economic systems. He wrote a pioneering paper on the essential analogy between statistical laws in physics and in social sciences (di Ettore Majorana (1942); Mantegna (2005, 2006)). However, during the following decades, only few physicists like Kadanoff (1971) or Montroll and Badger (1974) had an explicit interest for research in social or economic systems. It was not until the 1990's that physicists started turning to this interdisciplinary subject, and in the past years, they have made many successful attempts to approach problems in various fields of social sciences (e.g. de Oliveira *et al.* (1999); Stauffer *et al.* (2006); Chakrabarti *et al.* (2006)). In particular, in Quantitative Economics and Finance, physics research has begun to be complementary to the most traditional approaches such as mathematical (stochastic) finance. These various investigations, based on methods imported from or also used in physics, are the subject of the present paper.

[a] Electronic mail: anirban.chakraborti@ecp.fr
[b] Electronic mail: ioane.muni-toke@ecp.fr
[c] Electronic mail: marco.patriarca@kbfi.ee
[d] Electronic mail: frederic.abergel@ecp.fr

A. Bridging Physics and Economics

Economics deals with how societies efficiently use their resources to produce valuable commodities and distribute them among different people or economic agents (Samuelson (1998); Keynes (1973)). It is a discipline related to almost everything around us, starting from the marketplace through the environment to the fate of nations. At first sight this may seem a very different situation from that of physics, whose birth as a well defined scientific theory is usually associated with the study of particular mechanical objects moving with negligible friction, such as falling bodies and planets. However, a deeper comparison shows many more analogies than differences. On a general level, both economics and physics deal with "everything around us", despite with different perspectives. On a practical level, the goals of both disciplines can be either purely theoretical in nature or strongly oriented toward the improvement of the quality of life. On a more technical side, analogies often become equivalences. Let us give here some examples.

Statistical mechanics has been defined as the

> "branch of physics that combines the principles and procedures of statistics with the laws of both classical and quantum mechanics, particularly with respect to the field of thermodynamics. It aims to predict and explain the measurable properties of macroscopic systems on the basis of the properties and behaviour of the microscopic constituents of those systems."[1]

The tools of statistical mechanics or statistical physics (Reif (1985); Pathria (1996); Landau (1965)), that include extracting the average properties of a macroscopic system from the microscopic dynamics of the systems, are believed to prove useful for an economic system. Indeed, even though it is difficult or almost impossible to write down the "microscopic equations of motion" for an economic system with all the interacting entities, economic systems may be investigated at various size scales. Therefore, the understanding of the global behaviour of economic systems seems to need concepts such as stochastic dynamics, correlation effects, self-organization, self-similarity and scaling, and for their application we do not have to go into the detailed "microscopic" description of the economic system.

Chaos theory has had some impact in Economics modelling, e.g. in the work by Brock and Hommes (1998) or Chiarella *et al.* (2006). The theory of disordered systems has also played a core role in Econophysics and study of "complex systems". The term "complex systems" was coined to cover the great variety of such systems which include examples from physics, chemistry, biology and also social sciences. The concepts and methods of statistical physics turned out to be extremely useful in application to these diverse complex systems including economic systems. Many complex systems in natural and social environments share the characteristics of competition among interacting agents for resources and their adaptation to dynamically changing environment (Parisi (1999); Arthur (1999)). Hence, the concept of disordered systems helps for instance to go beyond the concept of representative agent, an approach prevailing in much of (macro)economics and criticized by many economists (see e.g. Kirman (1992); Gallegati and Kirman (1999)). Minority games and their physical formulations have been exemplary.

Physics models have also helped bringing new theories explaining older observations in Economics. The Italian social economist Pareto investigated a century ago the wealth of individuals in a stable economy (Pareto (1897)) by modelling them with the distribution $P(> x) \sim x^{-\alpha}$, where $P(> x)$ is the number of people having income greater than or equal to x and α is an exponent (known now as the Pareto exponent) which he estimated to be 1.5. To explain such empirical findings, physicists have come up with some very elegant and intriguing kinetic exchange models in recent times, and we will review these developments in the companion article. Though the economic activities of the agents are driven by various considerations like "utility maximization", the eventual exchanges of money in any trade can be simply viewed as money/wealth conserving two-body scatterings, as in the entropy maximization based kinetic theory of gases. This qualitative analogy seems to be quite old and both economists and natural scientists have already noted it in various contexts (Saha *et al.* (1950)). Recently, an equivalence between the two maximization principles have been quantitatively established (Chakrabarti and Chakrabarti (2010)).

Let us discuss another example of the similarities of interests and tools in Physics and Economics. The frictionless systems which mark the early history of physics were soon recognized to be rare cases: not only at microscopic scale – where they obviously represent an exception due to the unavoidable interactions with the environment – but also at the macroscopic scale, where fluctuations of internal or external origin make a prediction of their time evolution impossible. Thus equilibrium and non-equilibrium statistical mechanics, the theory of stochastic processes, and the theory of chaos, became main tools for studying real systems as well as an important part of the theoretical framework of modern physics. Very interestingly, the same mathematical tools have presided at the growth of classic modelling in Economics and more particularly in modern Finance. Following the works of Mandelbrot, Fama of the 1960s, physicists from 1990 onwards have studied the fluctuation of prices and universalities in context of scaling theories, etc. These links open the way for the use of a physics approach in Fi-

[1] In Encyclopædia Britannica. Retrieved June 11, 2010, from Encyclopædia Britannica Online.

nance, complementary to the widespread mathematical one.

B. Econophysics and Finance

Mathematical finance has benefited a lot in the past thirty years from modern probability theory – Brownian motion, martingale theory, etc. Financial mathematicians are often proud to recall the most well-known source of the interactions between Mathematics and Finance: five years before Einstein's seminal work, the theory of the Brownian motion was first formulated by the French mathematician Bachelier in his doctoral thesis (Bachelier (1900); Boness (1967); Haberman and Sibbett (1995)), in which he used this model to describe price fluctuations at the Paris Bourse. Bachelier had even given a course as a "free professor" at the Sorbonne University with the title: "Probability calculus with applications to financial operations and analogies with certain questions from physics" (see the historical articles in Courtault *et al.* (2000); Taqqu (2001); Forfar (2002)).

Then Itō, following the works of Bachelier, Wiener, and Kolmogorov among many, formulated the presently known Itō calculus (Itō and McKean (1996)). The geometric Brownian motion, belonging to the class of Itō processes, later became an important ingredient of models in Economics (Osborne (1959); Samuelson (1965)), and in the well-known theory of option pricing (Black and Scholes (1973); Merton (1973)). In fact, stochastic calculus of diffusion processes combined with classical hypotheses in Economics led to the development of the *arbitrage pricing theory* (Duffie (1996), Follmer and Schied (2004)). The deregulation of financial markets at the end of the 1980's led to the exponential growth of the financial industry. Mathematical finance followed the trend: stochastic finance with diffusion processes and exponential growth of financial derivatives have had intertwined developments. Finally, this relationship was carved in stone when the Nobel prize was given to M.S. Scholes and R.C. Merton in 1997 (F. Black died in 1995) for their contribution to the theory of option pricing and their celebrated "Black-Scholes" formula.

However, this whole theory is closely linked to classical economics hypotheses and has not been grounded enough with empirical studies of financial time series. The Black-Scholes hypothesis of Gaussian log-returns of prices is in strong disagreement with empirical evidence. Mandelbrot (1960, 1963) was one of the firsts to observe a clear departure from Gaussian behaviour for these fluctuations. It is true that within the framework of stochastic finance and martingale modelling, more complex processes have been considered in order to take into account some empirical observations: jump processes (see e.g. Cont and Tankov (2004) for a textbook treatment) and stochastic volatility (e.g. Heston (1993); Gatheral (2006)) in particular. But recent events on financial markets and the succession of financial crashes (see e.g. Kindleberger and Aliber (2005) for a historical perspective) should lead scientists to re-think basic concepts of modelling. This is where Econophysics is expected to come to play. During the past decades, the financial landscape has been dramatically changing: deregulation of markets, growing complexity of products. On a technical point of view, the ever rising speed and decreasing costs of computational power and networks have lead to the emergence of huge databases that record all transactions and order book movements up to the millisecond. The availability of these data should lead to models that are better empirically founded. Statistical facts and empirical models will be reviewed in this article and its companion paper. The recent turmoil on financial markets and the 2008 crash seem to plead for new models and approaches. The Econophysics community thus has an important role to play in future financial market modelling, as suggested by contributions from Bouchaud (2008), Lux and Westerhoff (2009) or Farmer and Foley (2009).

C. A growing interdisciplinary field

The chronological development of Econophysics has been well covered in the book of Roehner (2002). Here it is worth mentioning a few landmarks. The first article on analysis of finance data which appeared in a physics journal was that of Mantegna (1991). The first conference in Econophysics was held in Budapest in 1997 and has been since followed by numerous schools, workshops and the regular series of meetings: APFA (Application of Physics to Financial Analysis), WEHIA (Workshop on Economic Heterogeneous Interacting Agents), and Econophys-Kolkata, amongst others. In the recent years the number of papers has increased dramatically; the community has grown rapidly and several new directions of research have opened. By now renowned physics journals like the Reviews of Modern Physics, Physical Review Letters, Physical Review E, Physica A, Europhysics Letters, European Physical Journal B, International Journal of Modern Physics C, etc. publish papers in this interdisciplinary area. Economics and mathematical finance journals, especially Quantitative Finance, receive contributions from many physicists. The interested reader can also follow the developments quite well from the preprint server (www.arxiv.org). In fact, recently a new section called quantitative finance has been added to it. One could also visit the web sites of the *Econophysics Forum* (www.unifr.ch/econophysics) and *Econophysics.Org* (www.econophysics.org). Previous texts addressing Econophysics issues, such as Bouchaud and Potters (2000); Mantegna and Stanley (2007); Gabaix (2009), may be complementary to the present review. The first textbook in Econophysics (Sinha *et al.* (2010)) is also in press.

D. Organization of the review

This article aims at reviewing recent empirical and theoretical developments that use tools from Physics in the fields of Economics and Finance. In section II of this paper, empirical studies revealing statistical properties of financial time series are reviewed. We present the widely acknowledged "stylized facts" describing the distribution of the returns of financial assets. In section III we continue with the statistical properties observed on order books in financial markets. We reproduce most of the stated facts using our own high-frequency financial database. In the last part of this article (section IV), we review contributions on correlation on financial markets, among which the computation of correlations using high-frequency data, analyses based on random matrix theory and the use of correlations to build economics taxonomies. In the companion paper to follow, Econophysics models are reviewed through the point of view of agent-based modelling. Using previous work originally presented in the fields of behavioural finance and market microstructure theory, econophysicists have developed agent-based models of order-driven markets that are extensively reviewed there. We then turn to models of wealth distribution where an agent-based approach also prevails. As mentioned above, Econophysics models help bringing a new look on some Economics observations, and advances based on kinetic theory models are presented. Finally, a detailed review of game theory models and the now classic minority games composes the final part.

II. STATISTICS OF FINANCIAL TIME SERIES: PRICE, RETURNS, VOLUMES, VOLATILITY

Recording a sequence of prices of commodities or assets produce what is called time series. Analysis of financial time series has been of great interest not only to the practitioners (an empirical discipline) but also to the theoreticians for making inferences and predictions. The inherent uncertainty in the financial time series and its theory makes it specially interesting to economists, statisticians and physicists (Tsay (2005)).

Different kinds of financial time series have been recorded and studied for decades, but the scale changed twenty years ago. The computerization of stock exchanges that took place all over the world in the mid 1980's and early 1990's has lead to the explosion of the amount of data recorded. Nowadays, all transactions on a financial market are recorded *tick-by-tick*, i.e. every event on a stock is recorded with a timestamp defined up to the millisecond, leading to huge amounts of data. For example, as of today (2010), the Reuters Datascope Tick History (RDTH) database records roughly 25 gigabytes of data *every trading day*.

Prior to this improvement in recording market activity, statistics could be computed with daily data at best. Now scientists can compute intraday statistics in high-frequency. This allows to check known properties at new time scales (see e.g. section II B below), but also implies special care in the treatment (see e.g. the computation of correlation on high-frequency in section IV A below).

It is a formidable task to make an exhaustive review on this topic but we try to give a flavour of some of the aspects in this section.

A. "Stylized facts" of financial time series

The concept of "stylized facts" was introduced in macroeconomics around 1960 by Kaldor (1961), who advocated that a scientist studying a phenomenon "should be free to start off with a stylized view of the facts". In his work, Kaldor isolated several statistical facts characterizing macroeconomic growth over long periods and in several countries, and took these robust patterns as a starting point for theoretical modelling.

This expression has thus been adopted to describe empirical facts that arose in statistical studies of financial time series and that seem to be persistent across various time periods, places, markets, assets, etc. One can find many different lists of these facts in several reviews (e.g. Bollerslev *et al.* (1994); Pagan (1996); Guillaume *et al.* (1997); Cont (2001)). We choose in this article to present a minimum set of facts now widely acknowledged, at least for the prices of equities.

1. Fat-tailed empirical distribution of returns

Let p_t be the price of a financial asset at time t. We define its return over a period of time τ to be:

$$r_\tau(t) = \frac{p(t+\tau) - p(t)}{p(t)} \approx \log(p(t+\tau)) - \log(p(t)) \quad (1)$$

It has been largely observed – starting with Mandelbrot (1963), see e.g. Gopikrishnan *et al.* (1999) for tests on more recent data – and it is the first stylized fact, that the empirical distributions of financial returns and log-returns are fat-tailed. On figure 1 we reproduce the empirical density function of normalized log-returns from Gopikrishnan *et al.* (1999) computed on the S&P500 index. In addition, we plot similar distributions for unnormalized returns on a liquid French stock (BNP Paribas) with $\tau = 5$ minutes. This graph is computed by sampling a set of tick-by-tick data from 9:05am till 5:20pm between January 1st, 2007 and May 30th, 2008, i.e. 356 days of trading. Except where mentioned otherwise in captions, this data set will be used for all empirical graphs in this section. On figure 2, cumulative distribution in log-log scale from Gopikrishnan *et al.* (1999) is reproduced. We also show the same distribution in linear-log scale computed on our data for a larger time scale $\tau = 1$ day, showing similar behaviour.

Many studies obtain similar observations on different sets of data. For example, using two years of data on

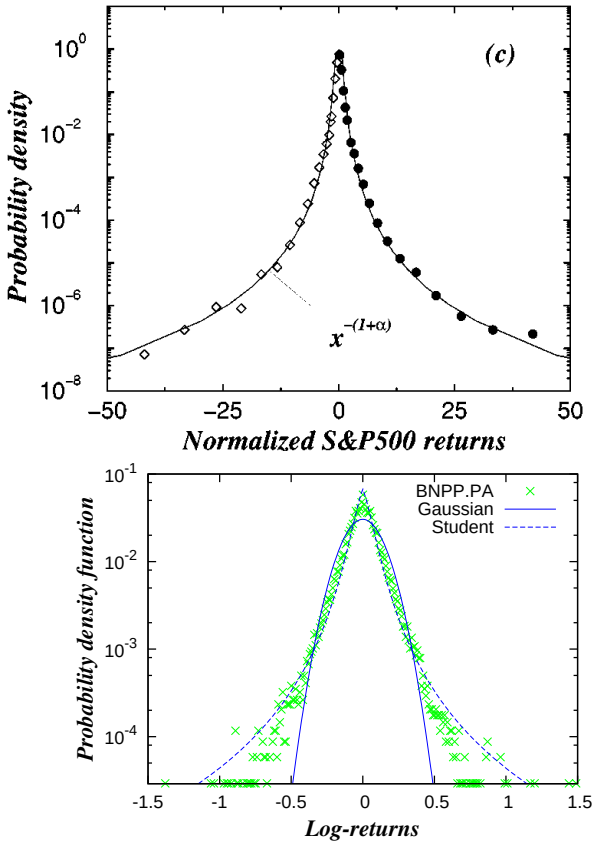

FIG. 1. (Top) Empirical probability density function of the normalized 1-minute S&P500 returns between 1984 and 1996. Reproduced from Gopikrishnan *et al.* (1999). (Bottom) Empirical probability density function of BNP Paribas unnormalized log-returns over a period of time $\tau = 5$ minutes.

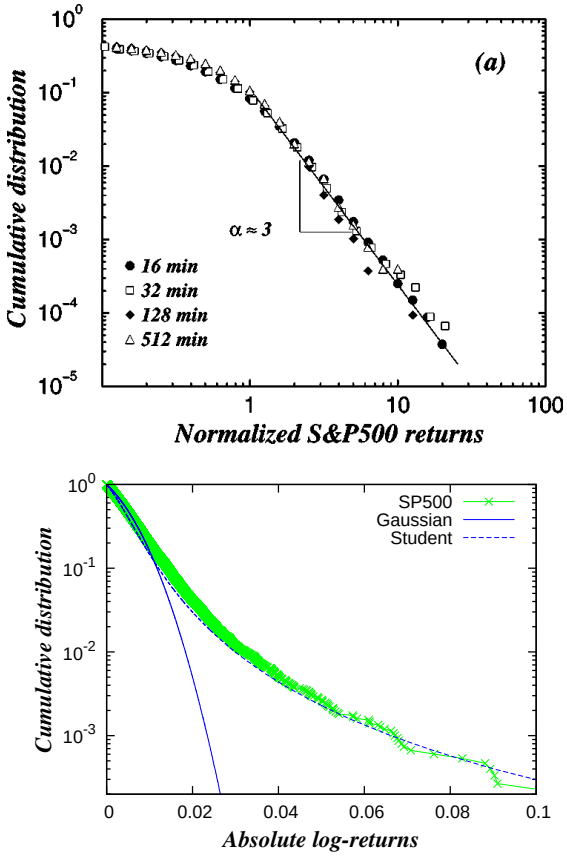

FIG. 2. Empirical cumulative distributions of S&P 500 daily returns. (Top) Reproduced from Gopikrishnan *et al.* (1999), in log-log scale. (Bottom) Computed using official daily close price between January 1st, 1950 and June 15th, 2009, i.e. 14956 values, in linear-log scale.

more than a thousand US stocks, Gopikrishnan *et al.* (1998) finds that the cumulative distribution of returns asymptotically follow a power law $F(r_\tau) \sim |r|^{-\alpha}$ with $\alpha > 2$ ($\alpha \approx 2.8 - 3$). With $\alpha > 2$, the second moment (the variance) is well-defined, excluding stable laws with infinite variance. There has been various suggestions for the form of the distribution: Student's-t, hyperbolic, normal inverse Gaussian, exponentially truncated stable, and others, but no general consensus exists on the exact form of the tails. Although being the most widely acknowledged and the most elementary one, this stylized fact is not easily met by all financial modelling. Gabaix *et al.* (2006) or Wyart and Bouchaud (2007) recall that efficient market theory have difficulties in explaining fat tails. Lux and Sornette (2002) have shown that models known as "rational expectation bubbles", popular in economics, produced very fat-tailed distributions ($\alpha < 1$) that were in disagreement with the statistical evidence.

2. Absence of autocorrelations of returns

On figure 3, we plot the autocorrelation of log-returns defined as $\rho(T) \sim \langle r_\tau(t+T) r_\tau(t) \rangle$ with τ =1 minute and 5 minutes. We observe here, as it is widely known (see e.g. Pagan (1996); Cont *et al.* (1997)), that there is no evidence of correlation between successive returns, which is the second "stylized-fact". The autocorrelation function decays very rapidly to zero, even for a few lags of 1 minute.

3. Volatility clustering

The third "stylized-fact" that we present here is of primary importance. Absence of correlation between returns must no be mistaken for a property of independence and identical distribution: price fluctuations are not identically distributed and the properties of the distribution change with time.

In particular, absolute returns or squared returns exhibit a long-range slowly decaying auto correlation func-

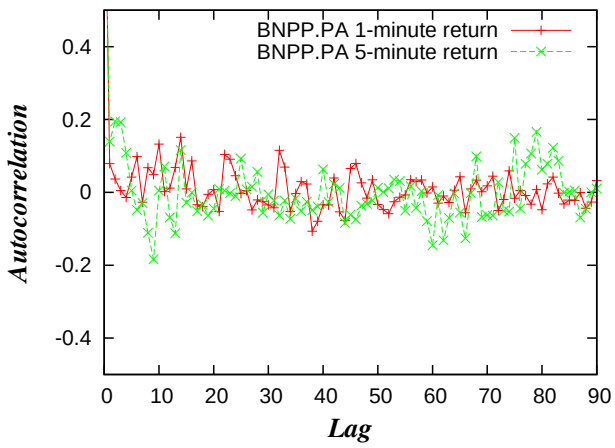

FIG. 3. Autocorrelation function of BNPP.PA returns.

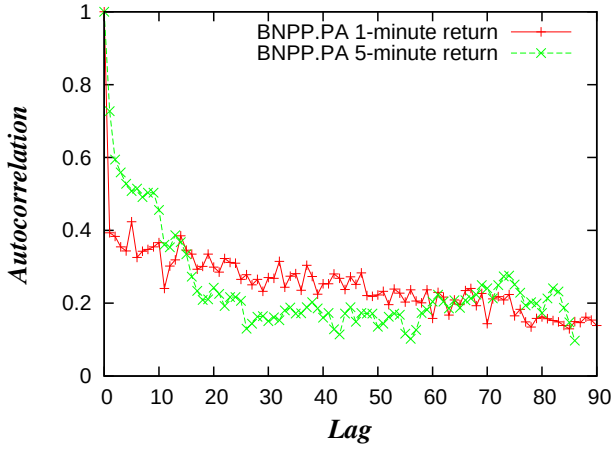

FIG. 4. Autocorrelation function of BNPP.PA absolute returns.

tion. This phenomena is widely known as "volatility clustering", and was formulated by Mandelbrot (1963) as "large changes tend to be followed by large changes – of either sign – and small changes tend to be followed by small changes".

On figure 4, the autocorrelation function of absolute returns is plotted for $\tau = 1$ minute and 5 minutes. The levels of autocorrelations at the first lags vary wildly with the parameter τ. On our data, it is found to be maximum (more than 70% at the first lag) for a returns sampled every five minutes. However, whatever the sampling frequency, autocorrelation is still above 10% after several hours of trading. On this data, we can grossly fit a power law decay with exponent 0.4. Other empirical tests report exponents between 0.1 and 0.3 (Cont *et al.* (1997); Liu *et al.* (1997); Cizeau *et al.* (1997)).

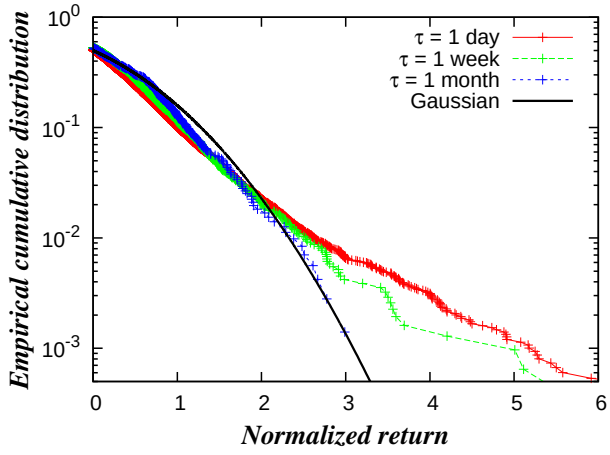

FIG. 5. Distribution of log-returns of S&P 500 daily, weekly and monthly returns. Same data set as figure 2 bottom.

4. Aggregational normality

It has been observed that as one increases the time scale over which the returns are calculated, the fat-tail property becomes less pronounced, and their distribution approaches the Gaussian form, which is the fourth "stylized-fact". This cross-over phenomenon is documented in Kullmann *et al.* (1999) where the evolution of the Pareto exponent of the distribution with the time scale is studied. On figure 5, we plot these standardized distributions for S&P 500 index between January 1st, 1950 and June 15th, 2009. It is clear that the larger the time scale increases, the more Gaussian the distribution is. The fact that the shape of the distribution changes with τ makes it clear that the random process underlying prices must have non-trivial temporal structure.

B. Getting the right "time"

1. Four ways to measure "time"

In the previous section, all "stylized facts" have been presented in *physical time, or calendar time*, i.e. time series were indexed, as we expect them to be, in hours, minutes, seconds, milliseconds. Let us recall here that tick-by-tick data available on financial markets all over the world is time-stamped up to the millisecond, but the order of magnitude of the guaranteed precision is much larger, usually one second or a few hundreds of milliseconds.

Calendar time is the time usually used to compute statistical properties of financial time series. This means that computing these statistics involves sampling, which might be a delicate thing to do when dealing for example with several stocks with different liquidity. Therefore, three other ways to keep track of time may be used.

Let us first introduce *event time*. Using this count, time is increased by one unit each time one order is submitted to the observed market. This framework is natural when dealing with the simulation of financial markets, as it will be showed in the companion paper. The main outcome of event time is its "smoothing" of data. In event time, intraday seasonality (lunch break) or outburst of activity consequent to some news are smoothed in the time series, since we always have one event per time unit.

Now, when dealing with time series of prices, another count of time might be relevant, and we call it *trade time* or *transaction time*. Using this count, time is increased by one unit each time a transaction happens. The advantage of this count is that limit orders submitted far away in the order book, and may thus be of lesser importance with respect to the price series, do not increase the clock by one unit.

Finally, going on with focusing on important events to increase the clock, we can use *tick time*. Using this count, time is increased by one unit each time the price changes. Thus consecutive market orders that progressively "eat" liquidity until the first best limit is removed in an order book are counted as one unit time.

Let us finish by noting that with these definitions, when dealing with mid prices, or bid and ask prices, a time series in event time can easily be extracted from a time series in calendar time. Furthermore, one can always extract a time series in trade time or in price time from a time series in event time. However, one cannot extract a series in price time from a series in trade time, as the latter ignores limit orders that are submitted inside the spread, and thus change mid, bid or ask prices without any transaction taking place.

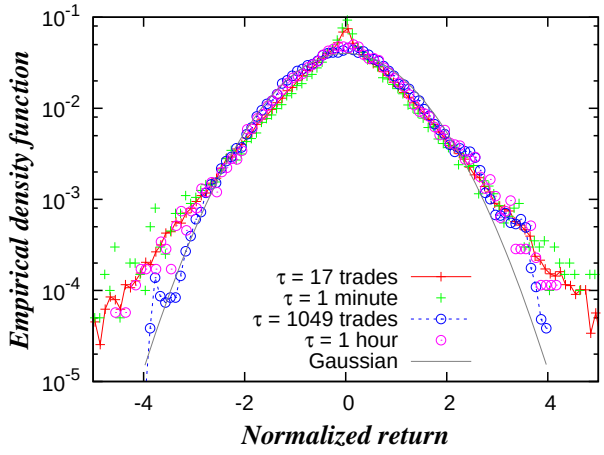

FIG. 6. Distribution of log-returns of stock BNPP.PA. This empirical distribution is computed using data from 2007, April 1st until 2008, May 31st.

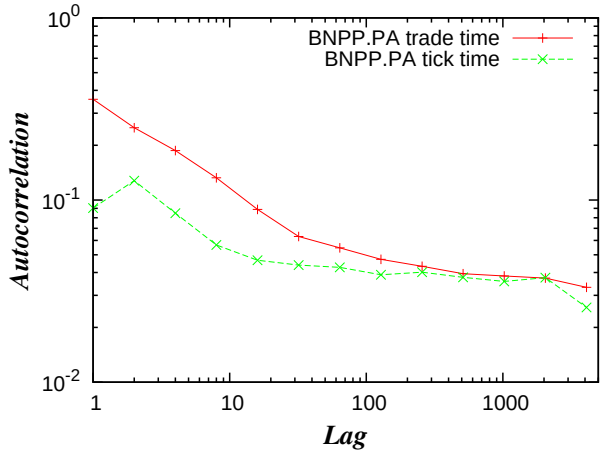

FIG. 7. Auto-correlation of trade signs for stock BNPP.PA.

2. Revisiting "stylized facts" with a new clock

Now, using the right clock might be of primary importance when dealing with statistical properties and estimators. For example, Griffin and Oomen (2008) investigates the standard realized variance estimator (see section IV A) in trade time and tick time. Muni Toke (2010) also recalls that the differences observed on a spread distribution in trade time and physical time are meaningful. In this section we compute some statistics complementary to the ones we have presented in the previous section II A and show the role of the clock in the studied properties.

a. Aggregational normality in trade time We have seen above that when the sampling size increases, the distribution of the log-returns tends to be more Gaussian. This property is much better seen using trade time. On figure 6, we plot the distributions of the log-returns for BNP Paribas stock using 2-month-long data in calendar time and trade time. Over this period, the average number of trade per day is 8562, so that 17 trades (resp. 1049 trades) corresponds to an average calendar time step of 1 minute (resp. 1 hour). We observe that the distribution of returns sampled every 1049 trades is much more Gaussian than the one sampled every 17 trades (aggregational normality), and that it is also more Gaussian that the one sampled every 1 hour (quicker convergence in trade time).

Note that this property appears to be valid in a multidimensional setting, see Huth and Abergel (2009).

b. Autocorrelation of trade signs in tick time It is well-known that the series of the signs of the trades on a given stock (usual convention: $+1$ for a transaction at the ask price, -1 for a transaction at the bid price) exhibit large autocorrelation. It has been observed in Lillo and Farmer (2004) for example that the autocorrelation function of the signs of trades (ϵ_n) was a slowly decaying function in $n^{-\alpha}$, with $\alpha \approx 0.5$. We compute this statistics for the trades on BNP Paribas stock from 2007, January 1st until 2008, May 31st. We plot the result in figure 7.

We find that the first values for short lags are about 0.3, and that the log-log plot clearly shows some power-law decay with roughly $\alpha \approx 0.7$.

A very plausible explanation of this phenomenon relies on the execution strategies of some major brokers on a given markets. These brokers have large transaction to execute on the account of some clients. In order to avoid market making move because of an inconsiderably large order (see below section III F on market impact), they tend to split large orders into small ones. We think that these strategies explain, at least partly, the large autocorrelation observed. Using data on markets where orders are publicly identified and linked to a given broker, it can be shown that the autocorrelation function of the order signs *of a given broker*, is even higher. See Bouchaud *et al.* (2009) for a review of these facts and some associated theories.

We present here another evidence supporting this explanation. We compute the autocorrelation function of order signs *in tick time*, i.e. taking only into account transactions that make the price change. Results are plotted on figure 7. We find that the first values for short lags are about 0.10, which is much smaller than the values observed with the previous time series. This supports the idea that many small transactions progressively "eat" the available liquidity at the best quotes. Note however that even in tick time, the correlation remains positive for large lags also.

3. Correlation between volume and volatility

Investigating time series of cotton prices, Clark (1973) noted that "trading volume and price change variance seem to have a curvilinear relationship". *Trade time* allows us to have a better view on this property: Plerou *et al.* (2000) and Silva and Yakovenko (2007) among others, show that the variance of log-returns after N trades, i.e. over a time period of N in trade time, is proprtional to N. We confirm this observation by plotting the second moment of the distribution of log-returns after N trades as a function of N for our data, as well as the average number of trades and the average volatility on a given time interval. The results are shown on figure 8 and 9.

This results are to be put in relation to the one presented in Gopikrishnan *et al.* (2000b), where the statistical properties of the number of shares traded $Q_{\Delta t}$ for a given stock in a fixed time interval Δt is studied. They analyzed transaction data for the largest 1000 stocks for the two-year period 1994-95, using a database that recorded every transaction for all securities in three major US stock markets. They found that the distribution $P(Q_{\Delta t})$ displayed a power-law decay as shown in Fig. 10, and that the time correlations in $Q_{\Delta t}$ displayed long-range persistence. Further, they investigated the relation between $Q_{\Delta t}$ and the number of transactions $N_{\Delta t}$ in a time interval Δt, and found that the long-range correlations in $Q_{\Delta t}$ were largely due to those of $N_{\Delta t}$.

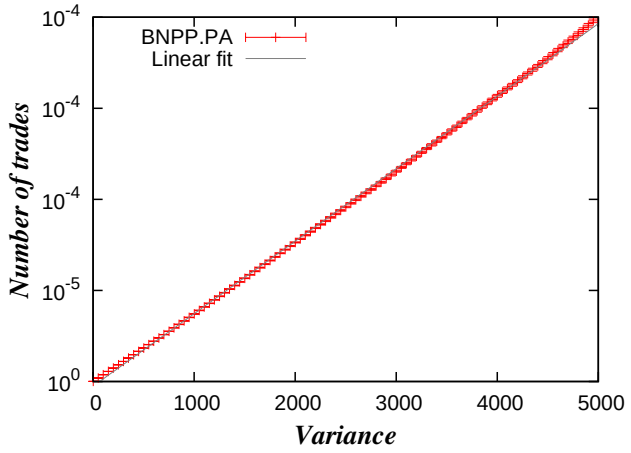

FIG. 8. Second moment of the distribution of returns over N trades for the stock BNPP.PA.

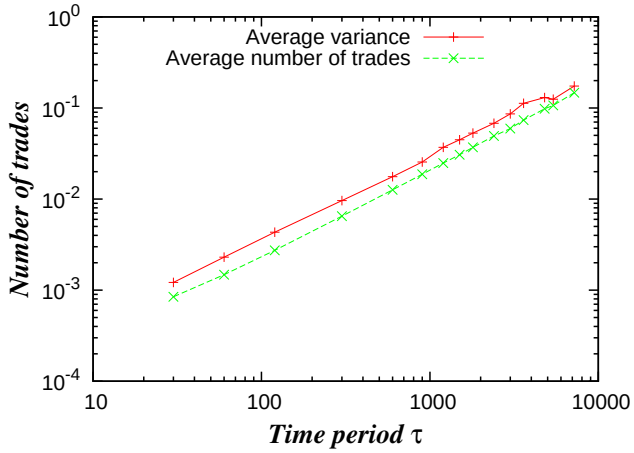

FIG. 9. Average number of trades and average volatility on a time period τ for the stock BNPP.PA.

Their results are consistent with the interpretation that the large equal-time correlation previously found between $Q_{\Delta t}$ and the absolute value of price change $|G_{\Delta t}|$ (related to volatility) were largely due to $N_{\Delta t}$.

Therefore, studying variance of price changer in *trade time* suggests that the number of trade is a good proxy for the unobserved volatility.

4. A link with stochastic processes: subordination

These empirical facts (aggregational normality in trade time, relationship between volume and volatility) reinforce the interest for models based on the subordination of stochastic processes, which had been introduced in financial modeling by Clark (1973).

Let us introduce it here. Assuming the proportionality between the variance $\langle x \rangle_\tau^2$ of the centred returns x and

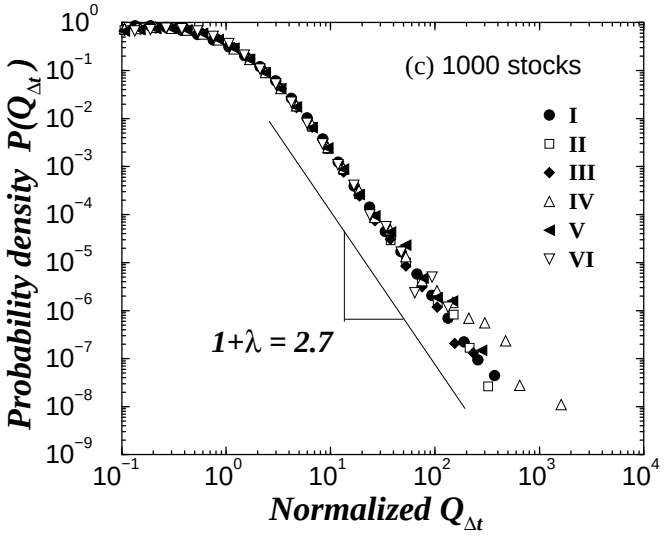

FIG. 10. Distribution of the number of shares traded $Q_{\Delta t}$. Adapted from Gopikrishnan *et al.* (2000b).

the number of trades N_τ over a time period τ, we can write:

$$\langle x \rangle_\tau^2 = \alpha N_\tau. \tag{2}$$

Therefore, assuming the normality in trade time, we can write the density function of log-returns after N trades as

$$f_N(x) = \frac{e^{\frac{-x^2}{2\alpha N}}}{\sqrt{2\pi\alpha N}}, \tag{3}$$

Finally, denoting $K_\tau(N)$ the probability density function of having N trades in a time period τ, the distribution of log returns in calendar time can be written

$$P_\tau(x) = \int_0^\infty \frac{e^{\frac{-x^2}{2\alpha N}}}{\sqrt{2\pi\alpha N}} K_\tau(N) dN. \tag{4}$$

This is the subordination of the Gaussian process x_N using the number of trades N_τ as the *directing process*, i.e. as the new clock. With this kind of modelization, it is expected, since P_N is gaussian, the observed non-gaussian behavior will come from $K_\tau(N)$. For example, some specific choice of directing processes may lead to a symmetric stable distribution (see Feller (1968)). Clark (1973) tests empirically a log-normal subordination with time series of prices of cotton. In a similar way, Silva and Yakovenko (2007) find that an exponential subordination with a kernel:

$$K_\tau(N) = \frac{1}{\eta\tau} e^{-\frac{N}{\eta\tau}}. \tag{5}$$

is in good agreement with empirical data. If the orders were submitted to the market in a independent way and at a constant rate η, then the distribution of the number of trade per time period τ should be a Poisson process with intensity $\eta\tau$. Therefore, the empirical fit of equation (5) is inconsistent with such a simplistic hypothesis of distribution of time of arrivals of orders. We will suggest in the next section some possible distributions that fit our empirical data.

III. STATISTICS OF ORDER BOOKS

The computerization of financial markets in the second half of the 1980's provided the empirical scientists with easier access to extensive data on order books. Biais *et al.* (1995) is an early study of the new data flows on the newly (at that time) computerized Paris Bourse. Variables crucial to a fine modeling of order flows and dynamics of order books are studied: time of arrival of orders, placement of orders, size of orders, shape of order book, etc. Many subsequent papers offer complementary empirical findings and modeling, e.g. Gopikrishnan *et al.* (2000a), Challet and Stinchcombe (2001), Maslov and Mills (2001), Bouchaud *et al.* (2002), Potters and Bouchaud (2003). Before going further in our review of available models, we try to summarize some of these empirical facts.

For each of the enumerated properties, we present new empirical plots. We use Reuters tick-by-tick data on the Paris Bourse. We select four stocks: France Telecom (FTE.PA) , BNP Paribas (BNPP.PA), Societe Générale (SOGN.PA) and Renault (RENA.PA). For any given stocks, the data displays time-stamps, traded quantities, traded prices, the first five best-bid limits and the first five best-ask limits. From now on, we will denote $a_i(t)$ (resp. $(b_j(t))$ the price of the i-th limit at ask (resp. j-th limit at bid). Except when mentioned otherwise, all statistics are computed using all trading days from Oct, 1st 2007 to May, 30th 2008, i.e. 168 trading days. On a given day, orders submitted between 9:05am and 5:20pm are taken into account, i.e. first and last minutes of each trading days are removed.

Note that we do not deal in this section with the correlations of the signs of trades, since statistical results on this fact have already been treated in section II B 2. Note also that although most of these facts are widely acknowledged, we will not describe them as new "stylized facts for order books" since their ranges of validity are still to be checked among various products/stocks, markets and epochs, and strong properties need to be properly extracted and formalized from these observations. However, we will keep them in mind as we go through the new trend of "empirical modeling" of order books.

Finally, let us recall that the markets we are dealing with are electronic order books with no official market maker, in which orders are submitted in a double auction and executions follow price/time priority. This type of exchange is now adopted nearly all over the world, but this was not obvious as long as computerization was not

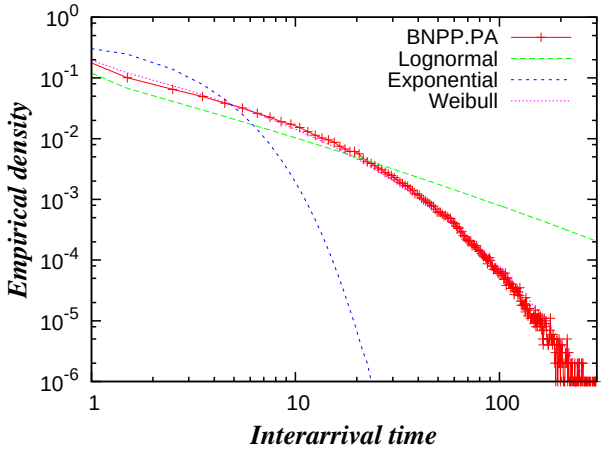

FIG. 11. Distribution of interarrival times for stock BNPP.PA in log-scale.

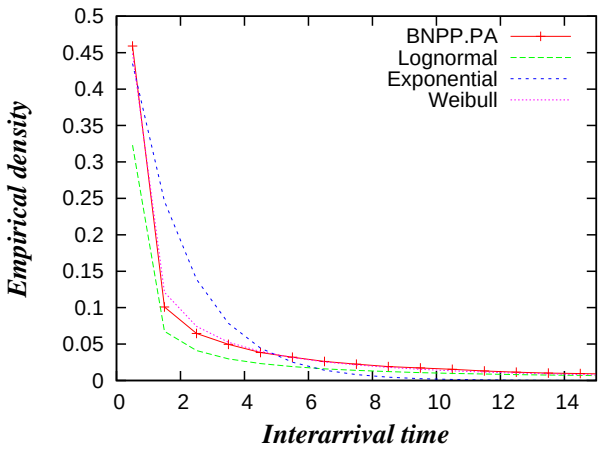

FIG. 12. Distribution of interarrival times for stock BNPP.PA (Main body, linear scale).

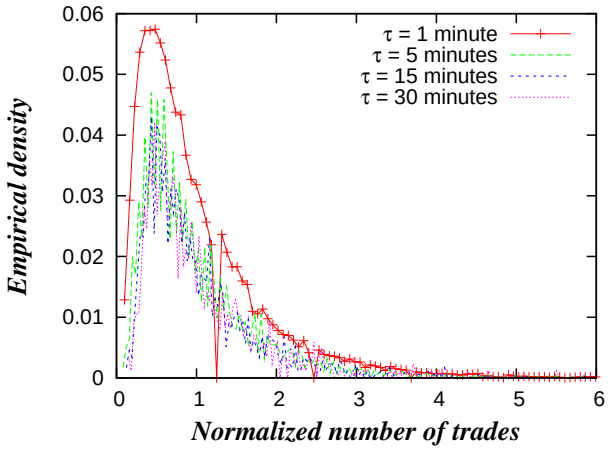

FIG. 13. Distribution of the number of trades in a given time period τ for stock BNPP.PA. This empirical distribution is computed using data from 2007, October 1st until 2008, May 31st.

complete. Different market mechanisms have been widely studied in the microstructure literature, see e.g. Garman (1976); Kyle (1985); Glosten (1994); O'Hara (1997); Biais *et al.* (1997); Hasbrouck (2007). We will not review this literature here (except Garman (1976) in our companion paper), as this would be too large a digression. However, such a literature is linked in many aspects to the problems reviewed in this paper.

A. Time of arrivals of orders

As explained in the previous section, the choice of the time count might be of prime importance when dealing with "stylized facts" of empirical financial time series. When reviewing the subordination of stochastic processes (Clark (1973); Silva and Yakovenko (2007)), we have seen that the Poisson hypothesis for the arrival times of orders is not empirically verified.

We compute the empirical distribution for interarrival times – or durations – of market orders on the stock BNP Paribas using our data set described in the previous section. The results are plotted in figures 11 and 12, both in linear and log scale. It is clearly observed that the exponential fit is not a good one. We check however that the Weibull distribution fit is potentially a very good one. Weibull distributions have been suggested for example in Ivanov *et al.* (2004). Politi and Scalas (2008) also obtain good fits with q-exponential distributions.

In the Econometrics literature, these observations of non-Poissonian arrival times have given rise to a large trend of modelling of irregular financial data. Engle and Russell (1997) and Engle (2000) have introduced autoregressive condition duration or intensity models that may help modelling these processes of orders' submission. See Hautsch (2004) for a textbook treatment.

Using the same data, we compute the empirical distribution of the number of transactions in a given time period τ. Results are plotted in figure 13. It seems that the log-normal and the gamma distributions are both good candidates, however none of them really describes the empirical result, suggesting a complex structure of arrival of orders. A similar result on Russian stocks was presented in Dremin and Leonidov (2005).

B. Volume of orders

Empirical studies show that the unconditional distribution of order size is very complex to characterize. Gopikrishnan *et al.* (2000a) and Maslov and Mills (2001) observe a power law decay with an exponent $1+\mu \approx 2.3-2.7$ for market orders and $1+\mu \approx 2.0$ for

limit orders. Challet and Stinchcombe (2001) emphasize on a clustering property: orders tend to have a "round" size in packages of shares, and clusters are observed around 100's and 1000's. As of today, no consensus emerges in proposed models, and it is plausible that such a distribution varies very wildly with products and markets.

In figure 14, we plot the distribution of volume of market orders for the four stocks composing our benchmark. Quantities are normalized by their mean. Power-law coefficient is estimated by a Hill estimator (see e.g. Hill (1975); de Haan *et al.* (2000)). We find a power law with exponent $1+\mu \approx 2.7$ which confirms studies previously cited. Figure 15 displays the same distribution for limit orders (of all available limits). We find an average value of $1+\mu \approx 2.1$, consistent with previous studies. However, we note that the power law is a poorer fit in the case of limit orders: data normalized by their mean collapse badly on a single curve, and computed coefficients vary with stocks.

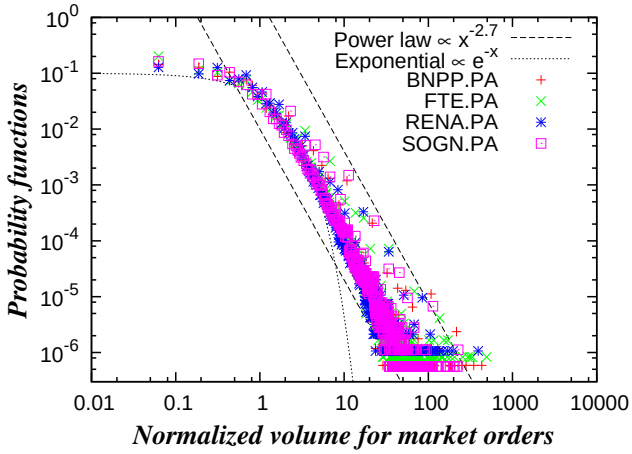

FIG. 14. Distribution of volumes of market orders. Quantities are normalized by their mean.

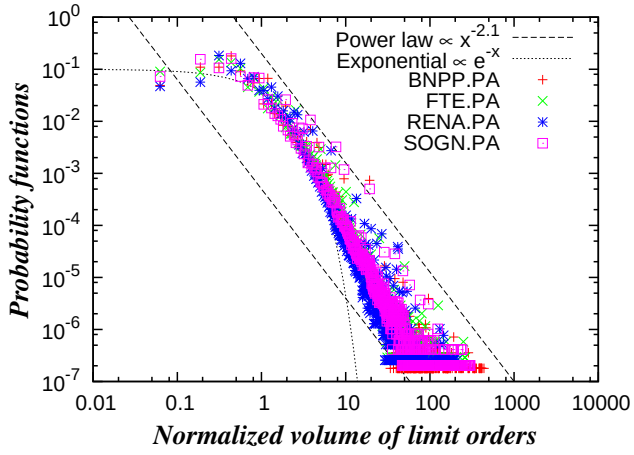

FIG. 15. Distribution of normalized volumes of limit orders. Quantities are normalized by their mean.

C. Placement of orders

a. Placement of arriving limit orders Bouchaud *et al.* (2002) observe a broad power-law placement around the best quotes on French stocks, confirmed in Potters and Bouchaud (2003) on US stocks. Observed exponents are quite stable across stocks, but exchange dependent: $1+\mu \approx 1.6$ on the Paris Bourse, $1+\mu \approx 2.0$ on the New York Stock Exchange, $1+\mu \approx 2.5$ on the London Stock Exchange. Mike and Farmer (2008) propose to fit the empirical distribution with a Student distribution with 1.3 degree of freedom.

We plot the distribution of the following quantity computed on our data set, i.e. using only the first five limits of the order book: $\Delta p = b_0(t-) - b(t)$ (resp. $a(t) - a_0(t-)$) if an bid (resp. ask) order arrives at price $b(t)$ (resp. $a(t)$), where $b_0(t-)$ (resp.$a_0(t-)$) is the best bid (resp. ask) before the arrival of this order. Results are plotted on figures 16 (in semilog scale) and 17 (in linear scale). These graphs being computed with incomplete data (five best limits), we do not observe a placement as broad as in Bouchaud *et al.* (2002). However, our data makes it clear that fat tails are observed. We also observe an asymmetry in the empirical distribution: the left side is less broad than the right side. Since the left side represent limit orders submitted *inside* the spread, this is expected. Thus, the empirical distribution of the placement of arriving limit orders is maximum at zero (same best quote). We then ask the question: How is it translated in terms of shape of the order book ?

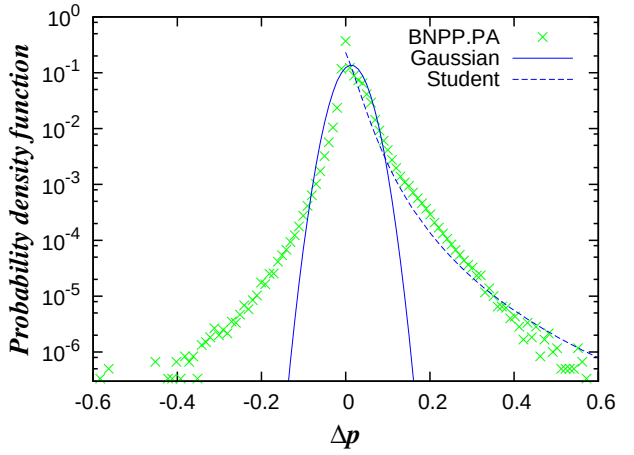

FIG. 16. Placement of limit orders using the same best quote reference in semilog scale. Data used for this computation is BNP Paribas order book from September 1st, 2007, until May 31st, 2008.

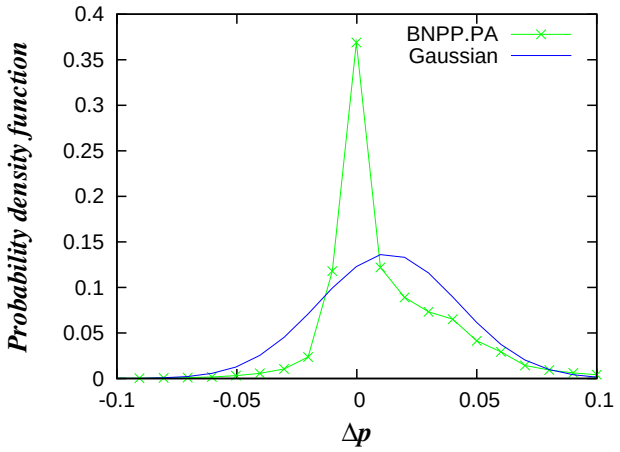

FIG. 17. Placement of limit orders using the same best quote reference in linear scale. Data used for this computation is BNP Paribas order book from September 1st, 2007, until May 31st, 2008.

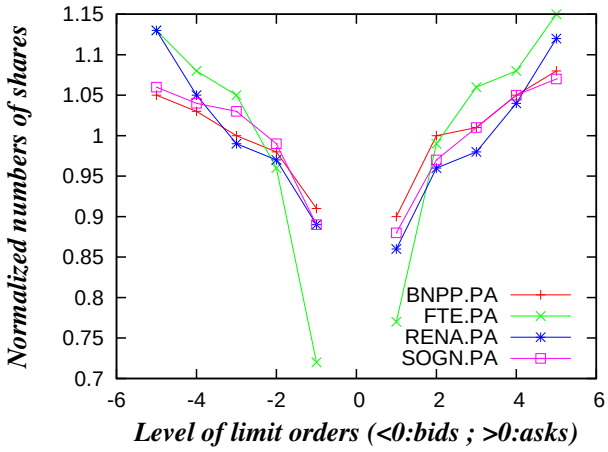

FIG. 18. Average quantity offered in the limit order book.

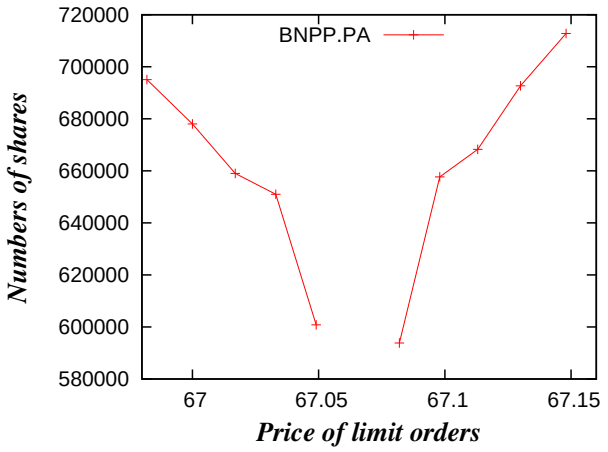

FIG. 19. Average limit order book: price and depth.

b. Average shape of the order book Contrary to what one might expect, it seems that the maximum of the average offered volume in an order book is located away from the best quotes (see e.g. Bouchaud *et al.* (2002)). Our data confirms this observation: the average quantity offered on the five best quotes grows with the level. This result is presented in figure 18. We also compute the average price of these levels in order to plot a cross-sectional graph similar to the ones presented in Biais *et al.* (1995). Our result is presented for stock BNP.PA in figure 19 and displays the expected shape. Results for other stocks are similar. We find that the average gap between two levels is constant among the five best bids and asks (less than one tick for FTE.PA, 1.5 tick for BNPP.PA, 2.0 ticks for SOGN.PA, 2.5 ticks for RENA.PA). We also find that the average spread is roughly twice as large the average gap (factor 1.5 for FTE.PA, 2 for BNPP.PA, 2.2 for SOGN.PA, 2.4 for RENA.PA).

D. Cancelation of orders

Challet and Stinchcombe (2001) show that the distribution of the average lifetime of limit orders fits a power law with exponent $1+\mu \approx 2.1$ for cancelled limit orders, and $1+\mu \approx 1.5$ for executed limit orders. Mike and Farmer (2008) find that in either case the exponential hypothesis (Poisson process) is not satisfied on the market.

We compute the average lifetime of cancelled and executed orders on our dataset. Since our data does not include a unique identifier of a given order, we reconstruct life time orders as follows: each time a cancellation is detected, we go back through the history of limit order submission and look for a matching order with same price and same quantity. If an order is not matched, we discard the cancellation from our lifetime data. Results are presented in figure 20 and 21. We observe a power law decay with coefficients $1+\mu \approx 1.3-1.6$ for both cancelled and executed limit orders, with little variations among stocks. These results are a bit different than the ones presented in previous studies: similar for executed limit orders, but our data exhibits a lower decay as for cancelled orders. Note that the observed cut-off in the distribution for lifetimes above 20000 seconds is due to the fact that we do not take into account execution or cancellation of orders submitted on a previous day.

E. Intraday seasonality

Activity on financial markets is of course not constant throughout the day. Figure 22 (resp. 23) plots the (normalized) number of market (resp. limit) orders arriving in a 5-minute interval. It is clear that a U-shape is ob-

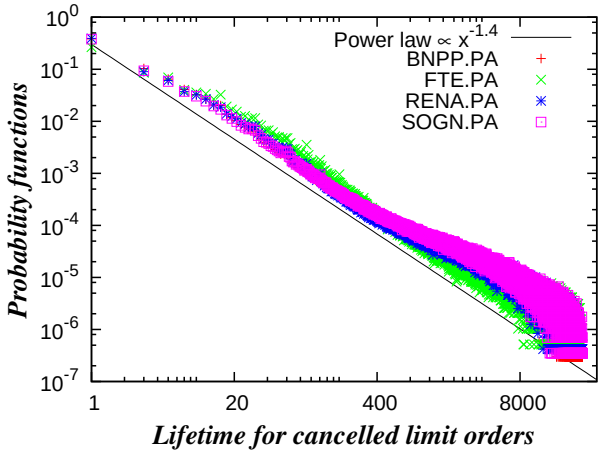

FIG. 20. Distribution of estimated lifetime of cancelled limit orders.

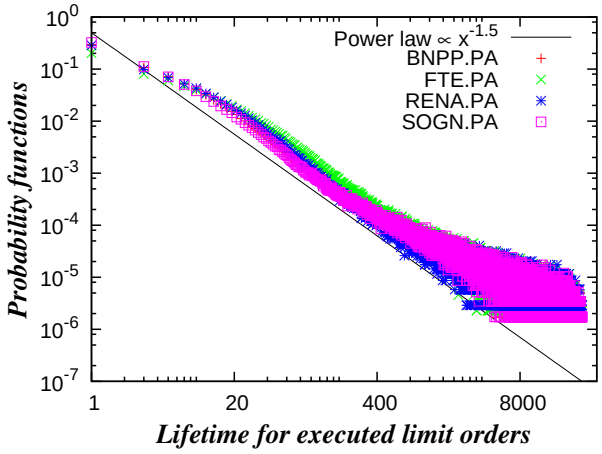

FIG. 21. Distribution of estimated lifetime of executed limit orders.

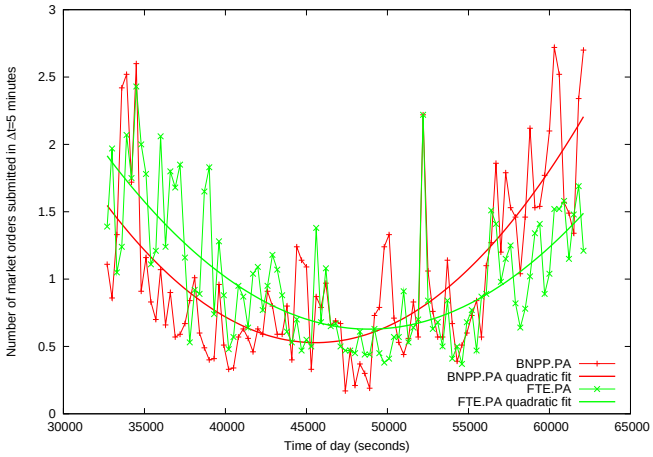

FIG. 22. Normalized average number of market orders in a 5-minute interval.

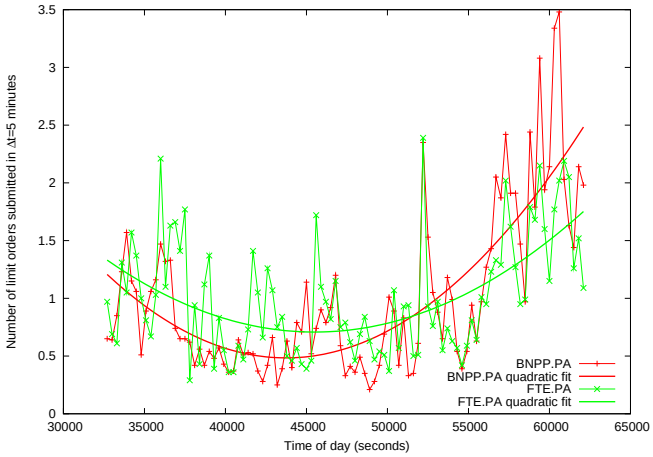

FIG. 23. Normalized average number of limit orders in a 5-minute interval.

served (an ordinary least-square quadratic fit is plotted): the observed market activity is larger at the beginning and the end of the day, and more quiet around mid-day. Such a U-shaped curve is well-known, see Biais *et al.* (1995), for example. On our data, we observe that the number of orders on a 5-minute interval can vary with a factor 10 throughout the day.

Challet and Stinchcombe (2001) note that the average number of orders submitted to the market in a period ΔT vary wildly during the day. The authors also observe that these quantities for market orders and limit orders are highly correlated. Such a type of intraday variation of the global market activity is a well-known fact, already observed in Biais *et al.* (1995), for example.

F. Market impact

The statistics we have presented may help to understand a phenomenon of primary importance for any financial market practitioner: the market impact, i.e. the relationship between the volume traded and the expected price shift once the order has been executed. On a first approximation, one understands that it is closely linked with many items described above: the volume of market orders submitted, the shape of the order book (how much pending limit orders are hit by one large market orders), the correlation of trade signs (one may assume that large orders are splitted in order to avoid a large market impact), etc.

Many empirical studies are available. An empirical study on the price impact of individual transactions on 1000 stocks on the NYSE is conducted in Lillo *et al.* (2003). It is found that proper rescaling make all the

curve collapse onto a single concave master curve. This function increases as a power that is the order of 1/2 for small volumes, but then increases more slowly for large volumes. They obtain similar results in each year for the period 1995 to 1998.

We will not review any further the large literature of market impact, but rather refer the reader to the recent exhaustive synthesis proposed in Bouchaud *et al.* (2009), where different types of impacts, as well as some theoretical models are discussed.

IV. CORRELATIONS OF ASSETS

The word "correlation" is defined as "a relation existing between phenomena or things or between mathematical or statistical variables which tend to vary, be associated, or occur together in a way not expected on the basis of chance alone"[2]. When we talk about correlations in stock prices, what we are really interested in are relations between variables such as stock prices, order signs, transaction volumes, etc. and more importantly how these relations affect the nature of the statistical distributions and laws which govern the price time series. This section deals with several topics concerning linear correlation observed in financial data. The first part deals with the important issue of computing correlations in high-frequency. As mentioned earlier, the computerization of financial exchanges has lead to the availability of huge amount of tick-by-tick data, and computing correlation using these intraday data raises lots of issues concerning usual estimators. The second and third parts deals with the use of correlation in order to cluster assets with potential applications in risk management problems.

A. Estimating covariance on high-frequency data

Let us assume that we observe d time series of prices or log-prices $p_i, i = 1, \ldots, d$, observed at times $t_m, m = 0, \ldots, M$. The usual estimator of the covariance of prices i and j is the *realized covariance estimator*, which is computed as:

$$\hat{\Sigma}_{ij}^{RV}(t) = \sum_{m=1}^{M} (p_i(t_m) - p_i(t_{m-1}))(p_j(t_m) - p_j(t_{m-1})). \tag{6}$$

The problem is that high-frequency tick-by-tick data record changes of prices when they happen, i.e. at random times. Tick-by-tick data is thus asynchronous, contrary to daily close prices for example, that are recorded at the same time for all the assets on a given exchange. Using standard estimators without caution, could be one cause for the "Epps effect", first observed in Epps (1979), which stated that "[c]orrelations among price changes in common stocks of companies in one industry are found to decrease with the length of the interval for which the price changes are measured." This has largely been verified since, e.g. in Bonanno *et al.* (2001) or Reno (2003). Hayashi and Yoshida (2005) shows that non-synchronicity of tick-by-tick data and necessary sampling of time series in order to compute the usual realized covariance estimator partially explain this phenomenon. We very briefly review here two covariance estimators that do not need any synchronicity (hence, sampling) in order to be computed.

1. The Fourier estimator

The Fourier estimator has been introduced by Malliavin and Mancino (2002). Let us assume that we have d time series of log-prices that are observations of Brownian semi-martingales p_i:

$$dp_i = \sum_{j=1}^{K} \sigma_{ij} dW_j + \mu_i dt, i = 1, \ldots, d. \tag{7}$$

The coefficient of the covariance matrix are then written $\Sigma_{ij}(t) = \sum_{k=1}^{K} \sigma_{ik}(t)\sigma_{jk}(t)$. Malliavin and Mancino (2002) show that the Fourier coefficient of $\Sigma_{ij}(t)$ are, with n_0 a given integer:

$$a_k(\Sigma_{ij}) = \lim_{N\to\infty} \frac{\pi}{N+1-n_0} \sum_{s=n_0}^{N} \frac{1}{2} \left[a_s(dp_i)a_{s+k}(dp_j) + b_{s+k}(dp_i)b_s(dp_j)\right], \tag{8}$$

$$b_k(\Sigma_{ij}) = \lim_{N\to\infty} \frac{\pi}{N+1-n_0} \sum_{s=n_0}^{N} \frac{1}{2} \left[a_s(dp_i)b_{s+k}(dp_j) - b_s(dp_i)a_{s+k}(dp_j)\right], \tag{9}$$

where the Fourier coefficients $a_k(dp_i)$ and $b_k(dp_i)$ of dp_i can be directly computed on the time series. Indeed, rescaling the time window on $[0, 2\pi]$ and using integration by parts, we have:

$$a_k(dp_i) = \frac{p(2\pi) - p(0)}{\pi} - \frac{k}{\pi} \int_0^{2\pi} \sin(kt) p_i(t) dt. \tag{10}$$

This last integral can be discretized and approximately computed using the times t_m^i of observations of the process p_i. Therefore, fixing a sufficiently large N, one can compute an estimator Σ_{ij}^F of the covariance of the processes i and j. See Reno (2003) or Iori and Precup (2007), for examples of empirical studies using this estimator.

2. The Hayashi-Yoshida estimator

Hayashi and Yoshida (2005) have proposed a simple estimator in order to compute covariance/correlation without any need for synchronicity of time series. As in the

[2] In Merriam-Webster Online Dictionary. Retrieved June 14, 2010, from http://www.merriam-webster.com/dictionary/correlations

Fourier estimator, it is assumed that the observed process is a Brownian semi-martingale. The time window of observation is easily partitioned into d family of intervals $\Pi^i = (U^i_m), i = 1, \ldots, d$, where $t^i_m = \inf\{U^i_{m+1}\}$ is the time of the m-th observation of the process i. Let us denote $\Delta p_i(U^i_m) = p_i(t^i_m) - p_i(t^i_{m-1})$. The *cumulative covariance estimator* as the authors named it, or the *Hayashi-Yoshida estimator* as it has been largely refered to, is then built as follows:

$$\hat{\Sigma}^{HY}_{ij}(t) = \sum_{m,n} \Delta p_i(U^i_m)\Delta p_j(U^j_n)\mathbf{1}_{\{U^i_m \cap U^j_n \neq \emptyset\}}. \qquad (11)$$

There is a large literature in Econometrics that tackles the new challenges posed by high-frequency data. We refer the reader, wishing to go beyond this brief presentation, to the econometrics reviews by Barndorff-Nielsen and Shephard (2007) or McAleer and Medeiros (2008), for example.

B. Correlation matrix and Random Matrix Theory

The stock market data being essentially a *multivariate* time series data, we construct correlation matrix to study its spectra and contrast it with the random multivariate data from coupled map lattice. It is known from previous studies that the empirical spectra of correlation matrices drawn from time series data, for most part, follow random matrix theory (RMT, see e.g. Gopikrishnan *et al.* (2001)).

1. Correlation matrix and Eigenvalue density

a. Correlation matrix If there are N assets with price $P_i(t)$ for asset i at time t, then the logarithmic return of stock i is $r_i(t) = \ln P_i(t) - \ln P_i(t-1)$, which for a certain consecutive sequence of trading days forms the return vector r_i. In order to characterize the synchronous time evolution of stocks, the equal time correlation coefficients between stocks i and j is defined as

$$\rho_{ij} = \frac{\langle r_i r_j\rangle - \langle r_i\rangle\langle r_j\rangle}{\sqrt{[\langle r_i^2\rangle - \langle r_i\rangle^2][\langle r_j^2\rangle - \langle r_j\rangle^2]}}, \qquad (12)$$

where $\langle ...\rangle$ indicates a time average over the trading days included in the return vectors. These correlation coefficients form an $N \times N$ matrix with $-1 \leq \rho_{ij} \leq 1$. If $\rho_{ij} = 1$, the stock price changes are completely correlated; if $\rho_{ij} = 0$, the stock price changes are uncorrelated, and if $\rho_{ij} = -1$, then the stock price changes are completely anti-correlated.

b. Correlation matrix of spatio-temporal series from coupled map lattices Consider a time series of the form $z'(x,t)$, where $x = 1, 2, ...n$ and $t = 1, 2....p$ denote the discrete space and time, respectively. In this, the time series at every spatial point is treated as a different variable. We define the normalised variable as

$$z(x,t) = \frac{z'(x,t) - \langle z'(x)\rangle}{\sigma(x)}, \qquad (13)$$

where the brackets $\langle.\rangle$ represent temporal averages and $\sigma(x)$ the standard deviation of z' at position x. Then, the equal-time cross-correlation matrix that represents the spatial correlations can be written as

$$S_{x,x'} = \langle z(x,t)\, z(x',t)\rangle, \quad x, x' = 1, 2, \ldots, n. \qquad (14)$$

The correlation matrix is symmetric by construction. In addition, a large class of processes are translation invariant and the correlation matrix can contain that additional symmetry too. We will use this property for our correlation models in the context of coupled map lattice. In time series analysis, the averages $\langle.\rangle$ have to be replaced by estimates obtained from finite samples. As usual, we will use the maximum likelihood estimates, $\langle a(t)\rangle \approx \frac{1}{p}\sum_{t=1}^{p} a(t)$. These estimates contain statistical uncertainties, which disappears for $p \to \infty$. Ideally, one requires $p \gg n$ to have reasonably correct correlation estimates. See Chakraborti *et al.* (2007) for details of parameters.

c. Eigenvalue Density The interpretation of the spectra of empirical correlation matrices should be done carefully if one wants to be able to distinguish between system specific signatures and universal features. The former express themselves in the smoothed level density, whereas the latter usually are represented by the fluctuations on top of this smooth curve. In time series analysis, the matrix elements are not only prone to uncertainty such as measurement noise on the time series data, but also statistical fluctuations due to finite sample effects. When characterizing time series data in terms of random matrix theory, one is not interested in these trivial sources of fluctuations which are present on every data set, but one would like to identify the significant features which would be shared, in principle, by an "infinite" amount of data without measurement noise. The eigenfunctions of the correlation matrices constructed from such empirical time series carry the information contained in the original time series data in a "graded" manner and they also provide a compact representation for it. Thus, by applying an approach based on random matrix theory, one tries to identify non-random components of the correlation matrix spectra as deviations from random matrix theory predictions (Gopikrishnan *et al.* (2001)).

We will look at the eigenvalue density that has been studied in the context of applying random matrix theory methods to time series correlations. Let $\mathcal{N}(\lambda)$ be the integrated eigenvalue density which gives the number of eigenvalues less than a given value λ. Then, the eigenvalue or level density is given by $\rho(\lambda) = \frac{d\mathcal{N}(\lambda)}{d\lambda}$. This can be obtained assuming random correlation matrix and is found to be in good agreement with the empirical time series data from stock market fluctuations. From Random

Matrix Theory considerations, the eigenvalue density for random correlations is given by

$$\rho_{rmt}(\lambda) = \frac{Q}{2\pi\lambda}\sqrt{(\lambda_{max}-\lambda)(\lambda-\lambda_{min})}\,, \qquad (15)$$

where $Q = N/T$ is the ratio of the number of variables to the length of each time series. Here, λ_{max} and λ_{min}, representing the maximum and minimum eigenvalues of the random correlation matrix respectively, are given by $\lambda_{max,min} = 1 + 1/Q \pm 2\sqrt{1/Q}$. However, due to presence of correlations in the empirical correlation matrix, this eigenvalue density is often violated for a certain number of dominant eigenvalues. They often correspond to system specific information in the data. In Fig. 24 we show the eigenvalue density for S&P500 data and also for the chaotic data from coupled map lattice. Clearly, both curves are qualitatively different. Thus, presence or absence of correlations in data is manifest in the spectrum of the corresponding correlation matrices.

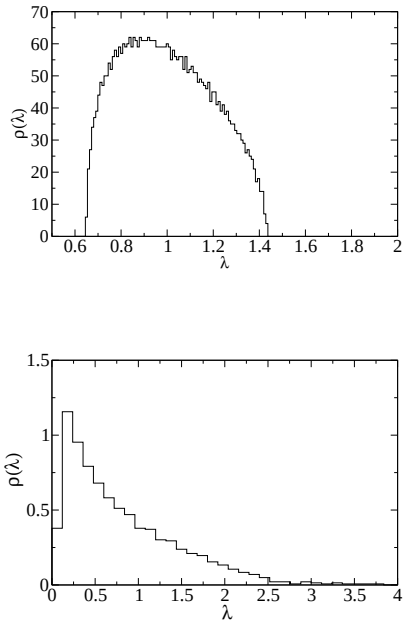

FIG. 24. The upper panel shows spectral density for multivariate spatio-temporal time series drawn from coupled map lattices. The lower panel shows the eigenvalue density for the return time series of the S&P500 stock market data (8938 time steps).

2. Earlier estimates and studies using Random Matrix Theory

Laloux *et al.* (1999) showed that results from the random matrix theory were useful to understand the statistical structure of the empirical correlation matrices appearing in the study of price fluctuations. The empirical

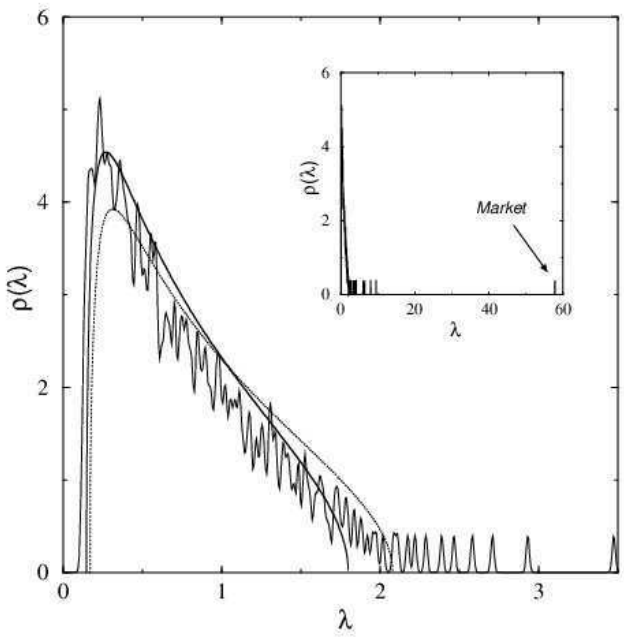

FIG. 25. Eigenvalue spectrum of the correlation matrices. Adapted from Laloux *et al.* (1999).

determination of a correlation matrix is a difficult task. If one considers N assets, the correlation matrix contains $N(N-1)/2$ mathematically independent elements, which must be determined from N time series of length T. If T is not very large compared to N, then generally the determination of the covariances is noisy, and therefore the empirical correlation matrix is to a large extent random. The smallest eigenvalues of the matrix are the most sensitive to this 'noise'. But the eigenvectors corresponding to these smallest eigenvalues determine the minimum risk portfolios in Markowitz theory. It is thus important to distinguish "signal" from "noise" or, in other words, to extract the eigenvectors and eigenvalues of the correlation matrix containing real information (those important for risk control), from those which do not contain any useful information and are unstable in time. It is useful to compare the properties of an empirical correlation matrix to a "null hypothesis"— a random matrix which arises for example from a finite time series of strictly uncorrelated assets. Deviations from the random matrix case might then suggest the presence of true information. The main result of their study was the remarkable agreement between the theoretical prediction (based on the assumption that the correlation matrix is random) and empirical data concerning the density of eigenvalues (shown in Fig. 25) associated to the time series of the different stocks of the S&P 500 (or other stock markets). Cross-correlations in financial data were also studied by Plerou *et al.* (1999, 2002). They analysed cross-correlations between price fluctuations of different stocks using methods of RMT. Using two large

databases, they calculated cross-correlation matrices of returns constructed from (i) 30-min returns of 1000 US stocks for the 2-yr period 1994–95, (ii) 30-min returns of 881 US stocks for the 2-yr period 1996–97, and (iii) 1-day returns of 422 US stocks for the 35-yr period 1962–96. They also tested the statistics of the eigenvalues λ_i of cross-correlation matrices against a "null hypothesis". They found that a majority of the eigenvalues of the cross-correlation matrices were within the RMT bounds $[\lambda_{min}, \lambda_{max}]$, as defined above, for the eigenvalues of random correlation matrices. They also tested the eigenvalues of the cross-correlation matrices within the RMT bounds for universal properties of random matrices and found good agreement with the results for the Gaussian orthogonal ensemble (GOE) of random matrices — implying a large degree of randomness in the measured cross-correlation coefficients. Furthermore, they found that the distribution of eigenvector components for the eigenvectors corresponding to the eigenvalues outside the RMT bounds displayed systematic deviations from the RMT prediction and that these "deviating eigenvectors" were stable in time. They analysed the components of the deviating eigenvectors and found that the largest eigenvalue corresponded to an influence common to all stocks. Their analysis of the remaining deviating eigenvectors showed distinct groups, whose identities corresponded to conventionally-identified business sectors.

C. Analyses of correlations and economic taxonomy

1. Models and theoretical studies of financial correlations

Podobnik *et al.* (2000) studied how the presence of correlations in physical variables contributes to the form of probability distributions. They investigated a process with correlations in the variance generated by a Gaussian or a truncated Levy distribution. For both Gaussian and truncated Levy distributions, they found that due to the correlations in the variance, the process "dynamically" generated power-law tails in the distributions, whose exponents could be controlled through the way the correlations in the variance were introduced. For a truncated Levy distribution, the process could extend a truncated distribution beyond the *truncation cutoff*, leading to a crossover between a Levy stable power law and their "dynamically-generated" power law. It was also shown that the process could explain the crossover behavior observed in the S&P 500 stock index.

Noh (2000) proposed a model for correlations in stock markets in which the markets were composed of several groups, within which the stock price fluctuations were correlated. The spectral properties of empirical correlation matrices (Plerou *et al.* (1999); Laloux *et al.* (1999)) were studied in relation to this model and the connection between the spectral properties of the empirical correlation matrix and the structure of correlations in stock markets was established.

The correlation structure of extreme stock returns were studied by Cizeau *et al.* (2001). It has been commonly believed that the correlations between stock returns increased in high volatility periods. They investigated how much of these correlations could be explained within a simple non-Gaussian one-factor description with time independent correlations. Using surrogate data with the true market return as the dominant factor, it was shown that most of these correlations, measured by a variety of different indicators, could be accounted for. In particular, their one-factor model could explain the level and asymmetry of empirical exceeding correlations. However, more subtle effects required an extension of the one factor model, where the variance and skewness of the residuals also depended on the market return.

Burda *et al.* (2001) provided a statistical analysis of three S&P 500 covariances with evidence for raw tail distributions. They studied the stability of these tails against reshuffling for the S&P 500 data and showed that the covariance with the strongest tails was robust, with a spectral density in remarkable agreement with random Levy matrix theory. They also studied the inverse participation ratio for the three covariances. The strong localization observed at both ends of the spectral density was analogous to the localization exhibited in the random Levy matrix ensemble. They showed that the stocks with the largest scattering were the least susceptible to correlations and were the likely candidates for the localized states.

2. Analyses using graph theory and economic taxonomy

Mantegna (1999) introduced a method for finding a hierarchical arrangement of stocks traded in financial market, through studying the clustering of companies by using correlations of asset returns. With an appropriate metric – based on the earlier explained correlation matrix coefficients ρ_{ij}'s between all pairs of stocks i and j of the portfolio, computed in Eq. 12 by considering the synchronous time evolution of the difference of the logarithm of daily stock price – a fully connected graph was defined in which the nodes are companies, or stocks, and the "distances" between them were obtained from the corresponding correlation coefficients. The minimum spanning tree (MST) was generated from the graph by selecting the most important correlations and it was used to identify clusters of companies. The hierarchical tree of the sub-dominant ultrametric space associated with the graph provided information useful to investigate the number and nature of the common economic factors affecting the time evolution of logarithm of price of well defined groups of stocks. Several other attempts have been made to obtain clustering from the huge correlation matrix.

Bonanno *et al.* (2001) studied the high-frequency cross-correlation existing between pairs of stocks traded in a financial market in a set of 100 stocks traded in US equity

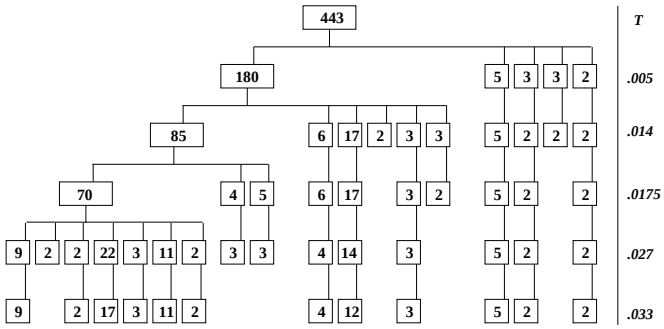

FIG. 26. The hierarchical structure of clusters of the S&P 500 companies in the ferromagnetic case. In the boxes the number of elements of the cluster are indicated. The clusters consisting of single companies are not indicated. Adapted from Kullmann *et al.* (2000).

markets. A hierarchical organization of the investigated stocks was obtained by determining a metric distance between stocks and by investigating the properties of the sub-dominant ultrametric associated with it. A clear modification of the hierarchical organization of the set of stocks investigated was detected when the time horizon used to determine stock returns was changed. The hierarchical location of stocks of the energy sector was investigated as a function of the time horizon. The hierarchical structure explored by the minimum spanning tree also seemed to give information about the influential power of the companies.

It also turned out that the hierarchical structure of the financial market could be identified in accordance with the results obtained by an independent clustering method, based on Potts super-paramagnetic transitions as studied by Kullmann *et al.* (2000), where the spins correspond to companies and the interactions are functions of the correlation coefficients determined from the time dependence of the companies' individual stock prices. The method is a generalization of the clustering algorithm by Blatt *et al.* (1996) to the case of anti-ferromagnetic interactions corresponding to anti-correlations. For the Dow Jones Industrial Average, no anti-correlations were observed in the investigated time period and the previous results obtained by different tools were well reproduced. For the S&P 500, where anti-correlations occur, repulsion between stocks modified the cluster structure of the $N = 443$ companies studied, as shown in Fig. 26. The efficiency of the method is represented by the fact that the figure matches well with the corresponding result obtained by the minimal spanning tree method, including the specific composition of the clusters. For example, at the lowest level of the hierarchy (highest temperature in the super-paramagnetic phase) the different industrial branches can be clearly identified: Oil, electricity, gold mining, etc. companies build separate clusters. The network of influence was investigated by means of a time-dependent correlation method by Kullmann *et al.* (2000). They studied the correlations as the function of the time shift between pairs of stock return time series of tick-by-tick data of the NYSE. They investigated whether any "pulling effect" between stocks existed or not, i.e. whether at any given time the return value of one stock influenced that of another stock at a different time or not. They found that, in general, two types of mechanisms generated significant correlation between any two given stocks. One was some kind of external effect (say, economic or political news) that influenced both stock prices simultaneously, and the change for both prices appeared at the same time, such that the maximum of the correlation was at zero time shift. The second effect was that, one of the companies had an influence on the other company indicating that one company's operation depended on the other, so that the price change of the influenced stock appeared latter because it required some time to react on the price change of the first stock displaying a "pulling effect". A weak but significant effect with the real data set was found, showing that in many cases the maximum correlation was at non-zero time shift indicating directions of influence between the companies, and the characteristic time was of the order of a few minutes, which was compatible with efficient market hypothesis. In the pulling effect, they found that in general, more important companies (which were traded more) pulled the relatively smaller companies.

The time dependent properties of the minimum spanning tree (introduced by Mantegna), called a 'dynamic asset tree', were studied by Onnela *et al.* (2003b). The nodes of the tree were identified with stocks and the distance between them was a unique function of the corresponding element of the correlation matrix. By using the concept of a central vertex, chosen as the most strongly connected node of the tree, the mean occupation layer was defined, which was an important characteristic of the tree. During crashes the strong global correlation in the market manifested itself by a low value of the mean occupation layer. The tree seemed to have a scale free structure where the scaling exponent of the degree distribution was different for 'business as usual' and 'crash' periods. The basic structure of the tree topology was very robust with respect to time. Let us discuss in more details how the dynamic asset tree was applied to studies of economic taxonomy.

a. Financial Correlation matrix and constructing Asset Trees Two different sets of financial data were used. The first set from the Standard & Poor's 500 index (S&P500) of the New York Stock Exchange (NYSE) from July 2, 1962 to December 31, 1997 contained 8939 daily closing values. The second set recorded the split-adjusted daily closure prices for a total of $N = 477$ stocks traded at the New York Stock Exchange (NYSE) over the period of 20 years, from 02-Jan-1980 to 31-Dec-1999. This amounted a total of 5056 prices per stock, indexed by time variable $\tau = 1, 2, \ldots, 5056$. For analysis and smoothing purposes, the data was divided time-wise into

M windows $t = 1, 2, ..., M$ of width T, where T corresponded to the number of daily returns included in the window. Note that several consecutive windows overlap with each other, the extent of which is dictated by the window step length parameter δT, which describes the displacement of the window and is also measured in trading days. The choice of window width is a trade-off between too noisy and too smoothed data for small and large window widths, respectively. The results presented here were calculated from monthly stepped four-year windows, i.e. $\delta T = 250/12 \approx 21$ days and $T = 1000$ days. A large scale of different values for both parameters were explored, and the cited values were found optimal(Onnela (2000)). With these choices, the overall number of windows is $M = 195$.

The earlier definition of correlation matrix, given by Eq. 12 is used. These correlation coefficients form an $N \times N$ correlation matrix $\mathbf{C}^t$, which serves as the basis for trees discussed below. An asset tree is then constructed according to the methodology by Mantegna (1999). For the purpose of constructing asset trees, a distance is defined between a pair of stocks. This distance is associated with the edge connecting the stocks and it is expected to reflect the level at which the stocks are correlated. A simple non-linear transformation $d_{ij}^t = \sqrt{2(1-\rho_{ij}^t)}$ is used to obtain distances with the property $2 \geq d_{ij} \geq 0$, forming an $N \times N$ symmetric distance matrix $\mathbf{D}^t$. So, if $d_{ij} = 0$, the stock price changes are completely correlated; if $d_{ij} = 2$, the stock price changes are completely anti-uncorrelated. The trees for different time windows are not independent of each other, but form a series through time. Consequently, this multitude of trees is interpreted as a sequence of evolutionary steps of a single *dynamic asset tree*. An additional hypothesis is required about the topology of the metric space: the ultrametricity hypothesis. In practice, it leads to determining the minimum spanning tree (MST) of the distances, denoted $\mathbf{T}^t$. The spanning tree is a simply connected acyclic (no cycles) graph that connects all N nodes (stocks) with $N-1$ edges such that the sum of all edge weights, $\sum_{d_{ij}^t \in \mathbf{T}^t} d_{ij}^t$, is minimum. We refer to the minimum spanning tree at time t by the notation $\mathbf{T}^t = (V, E^t)$, where V is a set of vertices and E^t is a corresponding set of unordered pairs of vertices, or edges. Since the spanning tree criterion requires all N nodes to be always present, the set of vertices V is time independent, which is why the time superscript has been dropped from notation. The set of edges E^t, however, does depend on time, as it is expected that edge lengths in the matrix $\mathbf{D}^t$ evolve over time, and thus different edges get selected in the tree at different times.

b. Market characterization We plot the distribution of (i) distance elements d_{ij}^t contained in the distance matrix $\mathbf{D}^t$ (Fig. 27), (ii) distance elements d_{ij} contained in the asset (minimum spanning) tree $\mathbf{T}^t$ (Fig. 28). In both plots, but most prominently in Fig. 27, there appears to be a discontinuity in the distribution between roughly 1986 and 1990. The part that has been cut out, pushed to the left and made flatter, is a manifestation of Black Monday (October 19, 1987), and its length along the time axis is related to the choice of window width T Onnela *et al.* (2003a,b). Also, note that in the distribution of tree edges in Fig. 28 most edges included in the tree seem to come from the area to the right of the value 1.1 in Fig. 27, and the largest distance element is $d_{max} = 1.3549$.

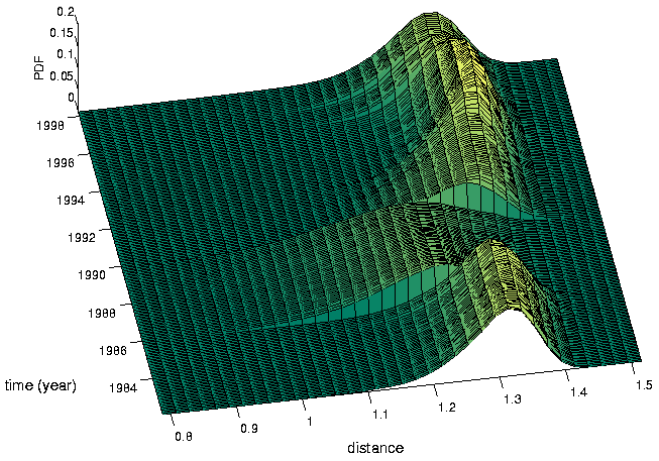

FIG. 27. Distribution of all $N(N-1)/2$ distance elements d_{ij} contained in the distance matrix $\mathbf{D}^t$ as a function of time.

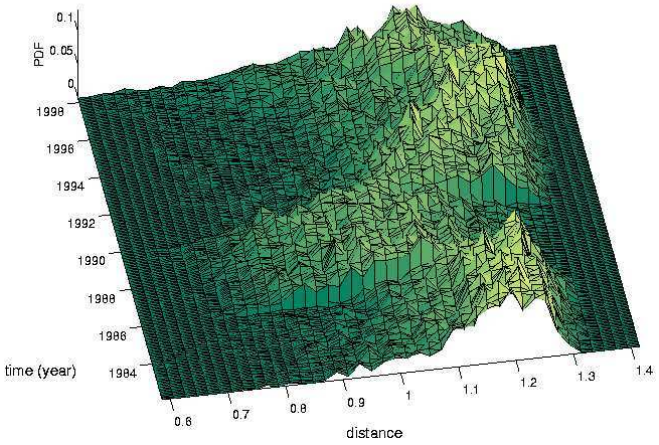

FIG. 28. Distribution of the $(N-1)$ distance elements d_{ij} contained in the asset (minimum spanning) tree $\mathbf{T}^t$ as a function of time.

Tree occupation and central vertex Let us focus on characterizing the spread of nodes on the tree, by introducing the quantity of *mean occupation layer*

$$l(t, v_c) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{lev}(v_i^t)\,, \tag{16}$$

where $\mathrm{lev}(v_i)$ denotes the level of vertex v_i. The levels, not to be confused with the distances d_{ij} between nodes, are measured in natural numbers in relation to the *central vertex* v_c, whose level is taken to be zero. Here the mean occupation layer indicates the layer on which the mass of the tree, on average, is conceived to be located. The central vertex is considered to be the parent of all other nodes in the tree, and is also known as the root of the tree. It is used as the *reference* point in the tree, against which the locations of all other nodes are relative. Thus all other nodes in the tree are children of the central vertex. Although there is an *arbitrariness* in the choice of the central vertex, it is proposed that the vertex is central, in the sense that any change in its price strongly affects the course of events in the market on the whole. Three alternative definitions for the central vertex were proposed in the studies, all yielding similar and, in most cases, identical outcomes. The idea is to find the node that is most strongly connected to its nearest neighbors. For example, according to one definition, the central node is the one with the highest *vertex degree*, i.e. the number of edges which are incident with (neighbor of) the vertex. Also, one may have either (i) static (fixed at all times) or (ii) dynamic (updated at each time step) central vertex, but again the results do not seem to vary significantly. The study of the variation of the topological properties and nature of the trees, with time were done.

Economic taxonomy Mantegna's idea of linking stocks in an ultrametric space was motivated *a posteriori* by the property of such a space to provide a meaningful economic taxonomy (Onnela *et al.* (2002)). Mantegna examined the meaningfulness of the taxonomy, by comparing the grouping of stocks in the tree with a third party reference grouping of stocks e.g. by their industry classifications (Mantegna (1999)). In this case, the reference was provided by Forbes (www.forbes.com), which uses its own classification system, assigning each stock with a sector (higher level) and industry (lower level) category. In order to visualize the grouping of stocks, a sample asset tree is constructed for a smaller dataset (shown in Fig. 29), which consists of 116 S&P 500 stocks, extending from the beginning of 1982 to the end of 2000, resulting in a total of 4787 price quotes per stock (Onnela *et al.* (2003b)). The window width was set at $T = 1000$, and the shown sample tree is located timewise at $t = t^*$, corresponding to 1.1.1998. The stocks in this dataset fall into 12 *sectors*, which are Basic Materials, Capital Goods, Conglomerates, Consumer/Cyclical, Consumer/Non-Cyclical, Energy, Financial, Healthcare, Services, Technology, Transportation and Utilities. The

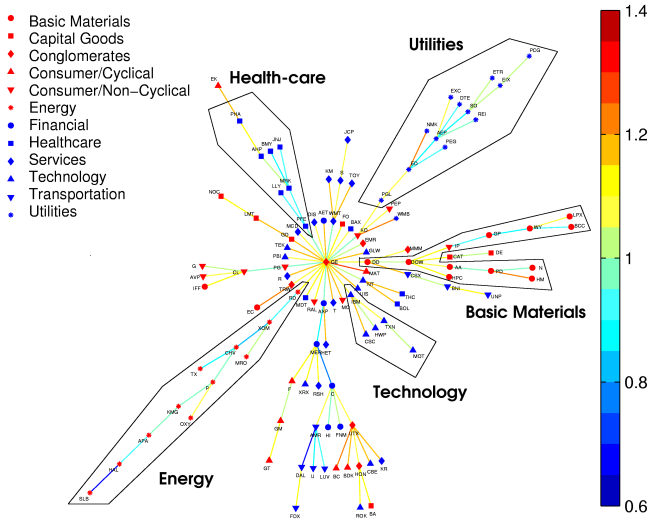

FIG. 29. Snapshot of a dynamic asset tree connecting the examined 116 stocks of the S&P 500 index. The tree was produced using four-year window width and it is centered on January 1, 1998. Business sectors are indicated according to Forbes (www.forbes.com). In this tree, General Electric (GE) was used as the central vertex and eight layers can be identified.

sectors are indicated in the tree (see Fig. 29) with different markers, while the industry classifications are omitted for reasons of clarity. The term sector is used exclusively to refer to the given third party classification system of stocks. The term *branch* refers to a subset of the tree, to all the nodes that share the specified common parent. In addition to the parent, it is needed to have a reference point to indicate the generational direction (i.e. who is who's parent) in order for a branch to be well defined. Without this reference there is absolutely no way to determine where one branch ends and the other begins. In this case, the reference is the central node. There are some branches in the tree, in which most of the stocks belong to just one sector, indicating that the branch is fairly homogeneous with respect to business sectors. This finding is in accordance with those of Mantegna (1999) , although there are branches that are fairly heterogeneous, such as the one extending directly downwards from the central vertex (see Fig. 29).

V. PARTIAL CONCLUSION

This first part of our review has shown statistical properties of financial data (time series of prices, order book structure, assets correlations). Some of these properties, such as fat tails of returns or volatility clustering, are widely known and acknowledged as "financial stylized facts". They are now largely cited in order to compare financial models, and reveal the lacks of many classical stochastic models of financial assets. Some other prop-

erties are newer findings that are obtained by studying high-frequency data of the whole order book structure. Volume of orders, interval time between orders, intraday seasonality, etc. are essential phenomenons to be understood when working in financial modelling. The important role of studies of correlations has been emphasized. Beside the technical challenges raised by high-frequency, many studies based for example on random matrix theory or clustering algorithms help getting a better grasp on some Economics problems. It is our belief that future modelling in finance will have to be partly based on Econophysics work on agent-based models in order to incorporate these "stylized facts" in a comprehensive way. Agent-based reasoning for order book models, wealth exchange models and game theoretic models will be reviewed in the following part of the review, to appear in a following companion paper.

ACKNOWLEDGEMENTS

The authors would like to thank their collaborators and two anonymous reviewers whose comments greatly helped improving the review. AC is grateful to B.K. Chakrabarti, K. Kaski, J. Kertesz, T. Lux, M. Marsili, D. Stauffer and V. Yakovenko for invaluable suggestions and criticisms.

Arthur, W., Complexity and the economy. *Science*, 1999, **284**, 107.

Bachelier, L., Theorie de la speculation. *Annales Scientifiques de l'Ecole Normale Superieure*, 1900, **III-17**, 21–86.

Barndorff-Nielsen, O.E. and Shephard, N., Econometric Society Monograph, Variation, jumps, market frictions and high frequency data in financial econometrics. In *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, 2007, Cambridge University Press.

Biais, B., Foucault, T. and Hillion, P., *Microstructure des marches financiers : Institutions, modeles et tests empiriques* 1997, Presses Universitaires de France - PUF.

Biais, B., Hillion, P. and Spatt, C., An empirical analysis of the limit order book and the order flow in the Paris Bourse. *Journal of Finance*, 1995, pp. 1655–1689.

Black, F. and Scholes, M., The pricing of options and corporate liabilities. *Journal of political economy*, 1973, **81**, 637.

Blatt, M., Wiseman, S. and Domany, E., Superparamagnetic Clustering of Data. *Physical Review Letters*, 1996, **76**, 3251.

Bollerslev, T., Engle, R.F. and Nelson, D.B., Chapter 49 Arch models. In *Handbook of Econometrics 4*, edited by R.F. Engle and D.L. McFadden, pp. 2959 – 3038, 1994, Elsevier.

Bonanno, G., Lillo, F. and Mantegna, R.N., High-frequency cross-correlation in a set of stocks. *Quantitative Finance*, 2001, **1**, 96.

Boness, A.J., In *The Random Character of Stock Market Prices*, edited by P.H. Cootner, chap. English translation of: L. Bachelier, Theorie de la Speculation, Annales scientifiques de l'Ecole Normale Superieure III-17, 1967, MIT Press.

Bouchaud, J.P., Mézard, M. and Potters, M., Statistical properties of stock order books: empirical results and models. *Quantitative Finance*, 2002, **2**, 251.

Bouchaud, J., Farmer, J.D., Lillo, F., Hens, T. and Schenk-Hopp, K.R., How Markets Slowly Digest Changes in Supply and Demand. In *Handbook of Financial Markets: Dynamics and Evolution*, pp. 57–160, 2009 (North-Holland: San Diego).

Bouchaud, J. and Potters, M., *Theory of Financial Risks: From Statistical Physics to Risk Management*, 2000, Cambridge University Press.

Bouchaud, J., Economics needs a scientific revolution. *Nature*, 2008, **455**, 1181.

Brock, W.A. and Hommes, C.H., Heterogeneous beliefs and routes to chaos in a simple asset pricing model. *Journal of Economic Dynamics and Control*, 1998, **22**, 1235–1274.

Burda, Z., Jurkiewicz, J., Nowak, M.A., Papp, G. and Zahed, I., Levy Matrices and Financial Covariances. *cond-mat/0103108*, 2001.

Chakrabarti, A.S. and Chakrabarti, B.K., Statistical Theories of Income and Wealth Distribution. *Economics E-Journal (open access)*, 2010, **4**.

Chakrabarti, B.K., Chakraborti, A. and Chatterjee, A. (Eds) *Econophysics and Sociophysics: Trends and Perspectives*, 1st 2006, Wiley - VCH, Berlin.

Chakraborti, A., Patriarca, M. and Santhanam, M.S., Financial Time-series Analysis: a Brief Overview. In *Econophysics of Markets and Business Networks*, 2007 (Springer: Milan).

Challet, D. and Stinchcombe, R., Analyzing and modeling 1+ 1d markets. *Physica A: Statistical Mechanics and its Applications*, 2001, **300**, 285–299.

Chiarella, C., He, X. and Hommes, C., A dynamic analysis of moving average rules. *Journal of Economic Dynamics and Control*, 2006, **30**, 1729–1753.

Cizeau, P., Potters, M. and Bouchaud, J., Correlation structure of extreme stock returns. *Quantitative Finance*, 2001, **1**, 217.

Cizeau, P., Liu, Y., Meyer, M., Peng, C.K. and Stanley, H.E., Volatility distribution in the S& P500 stock index. *Physica A: Statistical and Theoretical Physics*, 1997, **245**, 441 – 445.

Clark, P.K., A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices. *Econometrica*, 1973, **41**, 135–55.

Cont, R., Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 2001, **1**, 223–236.

Cont, R., Potters, M. and Bouchaud, J.P., Scale Invariance and Beyond. In *Proceedings of the Scale Invariance and Beyond: Les Houches Workshop 1997*, edited by F.G. B. Dubrulle and D. Sornette, 1997, Springer.

Cont, R. and Tankov, P., *Financial modelling with jump processes* 2004, Chapman & Hall/CRC.

Courtault, J., Kabanov, Y., Bru, B., Crepel, P., Lebon, I. and Le Marchand, A., Louis Bachelier on the centenary of Theorie de la Speculation. *Mathematical Finance*, 2000, **10**, 339–353.

de Haan, L., Resnick, S. and Drees, H., How to make a Hill plot. *The Annals of Statistics*, 2000, **28**, 254–274.

de Oliveira, S.M., de Oliveira, P.M.C. and Stauffer, D., *Evolution, Money, War and Computers* 1999, B. G. Teubner, Stuttgart-Leipzig.

di Ettore Majorana, N., Il valore delle leggi statistiche nella fisica e nelle scienze sociali. *Scientia*, 1942, **36**, 58–66.

Dremin, I. and Leonidov, A., On distribution of number of trades in different time windows in the stock market. *Physica A: Statistical Mechanics and its Applications*, 2005, **353**, 388 – 402.

Duffie, D., *Dynamic asset pricing theory* 1996, Princeton University Press Princeton, NJ.

Engle, R.F., The Econometrics of Ultra-High-Frequency Data. *Econometrica*, 2000, **68**, 1–22.

Engle, R.F. and Russell, J.R., Forecasting the frequency of changes in quoted foreign exchange prices with the autoregressive conditional duration model. *Journal of Empirical Finance*, 1997, **4**, 187–212.

Epps, T.W., Comovements in Stock Prices in the Very Short Run. *Journal of the American Statistical Association*, 1979, **74**, 291–298.

Farmer, J.D. and Foley, D., The economy needs agent-based modelling. *Nature*, 2009, **460**, 685–686.

Feller, W., *Introduction to the Theory of Probability and its Applications, Vol. 2* 1968, Wiley, New York.

Follmer, H. and Schied, A., *Stochastic Finance: An Introduction In Discrete Time 2*, 2nd Revised edition 2004, Walter de Gruyter & Co.

Forfar, D. Louis Bachelier, In: The MacTutor history of mathematics archive (published online), J. O'Connor and E. F. Robertson

eds. 2002.
Gabaix, X., Power Laws in Economics and Finance. *Annual Review of Economics*, 2009, **1**, 255–294.
Gabaix, X., Gopikrishnan, P., Plerou, V. and Stanley, H.E., Institutional Investors and Stock Market Volatility. *Quarterly Journal of Economics*, 2006, **121**, 461–504.
Gallegati, M. and Kirman, A.P. (Eds) *Beyond the Representative Agent*, 1st 1999, Edward Elgar Publishing.
Garman, M.B., Market microstructure. *Journal of Financial Economics*, 1976, **3**, 257 – 275.
Gatheral, J., *The volatility surface: a practitioner's guide* 2006, Wiley.
Glosten, L.R., Is the Electronic Open Limit Order Book Inevitable?. *The Journal of Finance*, 1994, **49**, 1127–1161.
Gopikrishnan, P., Plerou, V., Gabaix, X. and Stanley, H.E., Statistical properties of share volume traded in financial markets. *Physical Review - Series E-*, 2000a, **62**, 4493–4496.
Gopikrishnan, P., Meyer, M., Amaral, L.A. and Stanley, H.E., Inverse Cubic Law for the Probability Distribution of Stock Price Variations. *The European Physical Journal B*, 1998, **3**, 139.
Gopikrishnan, P., Plerou, V., Amaral, L.A., Meyer, M. and Stanley, H.E., Scaling of the distribution of fluctuations of financial market indices. *Phys. Rev. E*, 1999, **60**, 5305–5316.
Gopikrishnan, P., Plerou, V., Gabaix, X. and Stanley, H.E., Statistical properties of share volume traded in financial markets. *Physical Review E*, 2000b, **62**, R4493.
Gopikrishnan, P., Rosenow, B., Plerou, V. and Stanley, H.E., Quantifying and interpreting collective behavior in financial markets. *Physical Review E*, 2001, **64**, 035106.
Griffin, J.E. and Oomen, R.C.A., Sampling returns for realized variance calculations: tick time or transaction time?. *Econometric Reviews*, 2008, **27**, 230–253.
Guillaume, D., Dacorogna, M., Davé, R., Müller, U., Olsen, R. and Pictet, O., From the bird's eye to the microscope: A survey of new stylized facts of the intra-daily foreign exchange markets. *Finance and Stochastics*, 1997, **1**, 95–129.
Haberman, S. and Sibbett, T.A. (Eds), English translation of: Louis bachelier, Théorie de la spéculation, Annales scientifiques de l'Ecole Normale Superieure. In *History of Actuarial Science 7*, 1995, Pickering and Chatto Publishers, London.
Hasbrouck, J., *Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading* 2007, Oxford University Press, USA.
Hautsch, N., *Modelling irregularly spaced financial data* 2004, Springer.
Hayashi, T. and Yoshida, N., On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli*, 2005, **11**, 359–379.
Heston, S., A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Rev. Financ. Stud.*, 1993, **6**, 327–343.
Hill, B.M., A Simple General Approach to Inference About the Tail of a Distribution. *The Annals of Statistics*, 1975, **3**, 1163–1174.
Huth, N. and Abergel, F., The Times Change: Multivariate Subordination, Empirical Facts. *SSRN eLibrary*, 2009.
Iori, G. and Precup, O.V., Weighted network analysis of high-frequency cross-correlation measures. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 2007, **75**, 036110–7.
Itō, K. and McKean, H., *Diffusion Processes and Their Sample Paths* 1996 (Springer: Berlin).
Ivanov, P.C., Yuen, A., Podobnik, B. and Lee, Y., Common scaling patterns in intertrade times of U. S. stocks. *Phys. Rev. E*, 2004, **69**, 056107.
Kadanoff, L., From simulation model to public policy: An examination of Forrester's" Urban Dynamics". *Simulation*, 1971, **16**, 261.
Kaldor, N., Economic growth and capital accumulation. *The Theory of Capital, Macmillan, London*, 1961.
Keynes, J.M., *The general theory of employment, Interest and Money* 1973, The Royal Economic Society, Macmillan Press, London.
Kindleberger, C.P. and Aliber, R.Z., *Manias, Panics, And Crashes: A History Of Financial Crises*, Fifth edition 2005, John Wiley & Sons.
Kirman, A., Whom or what does the representative individual represent?. *The Journal of Economic Perspectives*, 1992, pp. 117–136.
Kullmann, L., Kertesz, J. and Mantegna, R.N., Identification of clusters of companies in stock indices via Potts super-paramagnetic transitions. *Physica A: Statistical Mechanics and its Applications*, 2000, **287**, 412419.
Kullmann, L., Toyli, J., Kertesz, J., Kanto, A. and Kaski, K., Characteristic times in stock market indices. *Physica A: Statistical Mechanics and its Applications*, 1999, **269**, 98 – 110.
Kyle, A.S., Continuous Auctions and Insider Trading. *Econometrica*, 1985, **53**, 1315–1335.
Laloux, L., Cizeau, P., Bouchaud, J. and Potters, M., Noise Dressing of Financial Correlation Matrices. *Physical Review Letters*, 1999, **83**, 1467.
Landau, L.D., *Statistical Physics, Vol. 5 of Theoretical Physics* 1965, Pergamon Press, Oxford.
Lillo, F., Farmer, D. and Mantegna, R., Econophysics: Master curve for price-impact function. *Nature*, 2003, **421**, 130, 129.
Lillo, F. and Farmer, J.D., The Long Memory of the Efficient Market. *Studies in Nonlinear Dynamics & Econometrics*, 2004, **8**.
Liu, Y., Cizeau, P., Meyer, M., Peng, C.K. and Stanley, H.E., Correlations in economic time series. *Physica A: Statistical and Theoretical Physics*, 1997, **245**, 437 – 440.
Lux, T. and Westerhoff, F., Economics crisis. *Nature Physics*, 2009, **5**, 2–3.
Lux, T. and Sornette, D., On Rational Bubbles and Fat Tails. *Journal of Money, Credit, and Banking*, 2002, **34**, 589–610.
Malliavin, P. and Mancino, M.E., Fourier series method for measurement of multivariate volatilities. *Finance and Stochastics*, 2002, **6**, 49–61.
Mandelbrot, B., The Pareto-Levy law and the distribution of income. *International Economic Review*, 1960, pp. 79–106.
Mandelbrot, B., The variation of certain speculative prices. *Journal of business*, 1963, **36**, 394.
Mantegna, R., Levy walks and enhanced diffusion in Milan stock exchange. *Physica. A*, 1991, **179**, 232–242.
Mantegna, R., Hierarchical structure in financial markets. *The European Physical Journal B - Condensed Matter and Complex Systems*, 1999, **11**, 193–197.
Mantegna, R., Presentation of the English translation of Ettore Majorana's paper: The value of statistical laws in physics and social sciences. *Quantitative Finance*, 2005, **5**, 133–140.
Mantegna, R., The Tenth Article of Ettore Majorana. *Europhysics News*, 2006, **37**.
Mantegna, R. and Stanley, H.E., *Introduction to Econophysics: Correlations and Complexity in Finance*, 2007, Cambridge University Press.
Maslov, S. and Mills, M., Price fluctuations from the order book perspective – empirical facts and a simple model. *Physica A: Statistical Mechanics and its Applications*, 2001, **299**, 234–246.
McAleer, M. and Medeiros, M.C., Realized volatility: A review. *Econometric Reviews*, 2008, **27**, 10–45.
Merton, R., Theory of rational option pricing. *The Bell Journal of Economics and Management Science*, 1973, pp. 141–183.
Mike, S. and Farmer, J.D., An empirical behavioral model of liquidity and volatility. *Journal of Economic Dynamics and Control*, 2008, **32**, 200–234.
Montroll, E. and Badger, W., *Introduction to quantitative aspects of social phenomena* 1974, Gordon and Breach New York.
Muni Toke, I., "Market making" in an order book model and its impact on the bid-ask spread. In *Econophysics of Order-Driven Markets*, 2010 (Springer: Milan).
Noh, J.D., Model for correlations in stock markets. *Physical Review E*, 2000, **61**, 5981.

O'Hara, M., *Market Microstructure Theory*, Second edition 1997, Blackwell Publishers.

Onnela, J.P., Chakraborti, A., Kaski, K. and Kertesz, J., Dynamic asset trees and Black Monday. *Physica A: Statistical Mechanics and its Applications*, 2003a, **324**, 247–252.

Onnela, J.P., Taxonomy of financial assets. Master's thesis, Helsinki University of Technology, Espoo, Finland 2000.

Onnela, J., Chakraborti, A., Kaski, K. and Kertesz, J., Dynamic asset trees and portfolio analysis. *The European Physical Journal B - Condensed Matter and Complex Systems*, 2002, **30**, 285–288.

Onnela, J., Chakraborti, A., Kaski, K., Kertesz, J. and Kanto, A., Dynamics of market correlations: Taxonomy and portfolio analysis. *Physical Review E*, 2003b, **68**, 056110.

Osborne, M.F.M., Brownian Motion in the Stock Market. *Operations Research*, 1959, **7**, 145–173.

Pagan, A., The econometrics of financial markets. *Journal of empirical finance*, 1996, **3**, 15–102.

Pareto, V., *Cours d'economie politique* 1897 (Rouge: Lausanne), Reprinted as a volume of Oeuvres Completes, G. Bousquet and G. Busino Eds., Droz, Genve, 1964.

Parisi, G., Complex systems: a physicist's viewpoint. *Physica A: Statistical Mechanics and its Applications*, 1999, **263**, 557 – 564 Proceedings of the 20th IUPAP International Conference on Statistical Physics.

Pathria, R.K., *Statistical Mechanics*, 2nd 1996, Butterworth-Heinemann, Oxford.

Plerou, V., Gopikrishnan, P., Nunes Amaral, L.A., Gabaix, X. and Eugene Stanley, H., Economic fluctuations and anomalous diffusion. *Phys. Rev. E*, 2000, **62**, R3023–R3026.

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N., Guhr, T. and Stanley, H.E., Random matrix approach to cross correlations in financial data. *Physical Review E*, 2002, **65**, 066126.

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N. and Stanley, H.E., Universal and Nonuniversal Properties of Cross Correlations in Financial Time Series. *Physical Review Letters*, 1999, **83**, 1471.

Podobnik, B., Ivanov, P.C., Lee, Y., Chessa, A. and Stanley, H.E., Systems with correlations in the variance: Generating power law tails in probability distributions. *EPL (Europhysics Letters)*, 2000, **50**, 711–717.

Politi, M. and Scalas, E., Fitting the empirical distribution of intertrade durations. *Physica A: Statistical Mechanics and its Applications*, 2008, **387**, 2025 – 2034.

Potters, M. and Bouchaud, J.P., More statistical properties of order books and price impact. *Physica A: Statistical Mechanics and its Applications*, 2003, **324**, 133–140.

Reif, F., *Fundamentals of Statistical and Thermal Physics* 1985, Mc Grow-Hill, Singapore.

Reno, R., A closer look at the Epps effect. *International Journal of Theoretical and Applied Finance*, 2003, **6**, 87–102.

Roehner, B., *Patterns of speculation: a study in observational econophysics* 2002, Cambridge University Press.

Saha, M.N., Srivastava, B.N. and Saha, M..S., *A treatise on heat* 1950 (Indian Press: Allahabad).

Samuelson, P., Proof that properly anticipated prices fluctuate randomly. *Management Review*, 1965, **6**.

Samuelson, P., *Economics* 1998, Mc Grow Hill, Auckland.

Silva, A.C. and Yakovenko, V.M., Stochastic volatility of financial markets as the fluctuating rate of trading: An empirical study. *Physica A: Statistical Mechanics and its Applications*, 2007, **382**, 278 – 285.

Sinha, S., Chatterjee, A., Chakraborti, A. and Chakrabarti, B.K., *Econophysics: An Introduction*, 1 2010, Wiley-VCH.

Stauffer, D., de Oliveira, S.M., de Oliveira, P.M.C. and de Sa Martins, J.S., *Biology, Sociology, Geology by Computational Physicists* 2006, Elsevier Science.

Taqqu, M., Bachelier and his times: A conversation with Bernard Bru. *Finance and Stochastics*, 2001, **5**, 3–32.

Tsay, R., *Analysis of financial time series* 2005, Wiley-Interscience.

Wyart, M. and Bouchaud, J.P., Self-referential behaviour, overreaction and conventions in financial markets. *Journal of Economic Behavior & Organization*, 2007, **63**, 1–24.